SCHOOL OF DATA SCIENCE AND FORECASTING

DEVI AHILYA VISHWAVIDYALAYA, INDORE (M.P)

Sustainalyze: Automated ESG Risk Analyzer with AI Insights

A Project Report

in partial fulfillment for the award of the degree

of

MASTER OF SCIENCE (M.Sc.)

in

DATA SCIENCE AND ANALYTICS

(Session: 2023-25)

Supervised by:

Dr. Vandit Hedau

Assistant Professor

Submitted by:

Nitishaw Saini

DS5B - 2322

Sustainalyze: Automated ESG Risk Analyzer with AI Insights

A Project Report

in partial fulfillment for the award of the degree

of

MASTER OF SCIENCE

in

DATA SCIENCE AND ANALYTICS

Submitted by

NITISHAW SAINI

DS5B-2306

EXTERNAL SUPERVISOR

INTERNAL SUPERVISOR

Dr. VANDIT HEDAU

H.O.D., School of Statistics

Associate Professor, SDSF

DAVV Indore

SCHOOL OF DATA SCIENCE AND FORECASTING

(UNIVERSITY TEACHING DEPARTMENT)

DEVI AHILYA VISHWAVIDYALAYA

Indore (M.P)

April, 2025

SCHOOL OF DATA SCIENCE AND FORECASTING
DEVI AHILYA VISHWAVIDYALAYA
INDORE (M.P)

STATEMENT OF ORIGINALITY

In accordance with the requirements for the Degree of Master of Science in DATA SCIENCE AND ANALYTICS, in SCHOOL OF DATA SCIENCE AND FORECASTING, I present this report entitled **"Sustainalyze: Automated ESG Risk Analyzer with AI Insights"**. This report is completed under the Supervision of:

EXTERNAL SUPERVISOR:

INTERNAL SUPERVISOR:

Dr. VANDIT HEDAU

H.O.D., SCHOOL OF STATISTICS

Associate Professor, School of Data Science & Forecasting

I declare that the work presented in the report is my own work except as acknowledged in the text and footnotes, and that to my knowledge this material has not been submitted either in whole or in part, for any other degree at this University or at any other such Institution

NITISHAW SAINI

28th April 2025

SCHOOL OF DATA SCIENCE AND FORECASTING

DEVI AHILYA VISHWAVIDYALAYA

INDORE (M.P)

RECOMMENDATION

This dissertation entitled "**Sustainalyze: Automated ESG Risk Analyzer with AI Insights**" submitted by **Nitishaw Saini** towards the partial fulfilment of Degree of Master of Science in Data Science and Analytics of Devi Ahilya Vishwavidyalaya, Indore is a satisfactory account of her project work and is recommended for the award of degree.

External Supervisor

Internal Supervisor
Dr. VANDIT HEDAU

Head of Department
SANJIV TOKEKAR SIR

SCHOOL OF DATA SCIENCE AND FORECASTING

DEVI AHILYA VISHWAVIDYALAYA

INDORE (M.P)

CERTIFICATE

This is to certify that Report entitled "**Sustainalyze: Automated ESG Risk Analyzer with AI Insights**" which is submitted by **NITISHAW SAINI** in partial fulfillment of the requirement for the award of degree **Master in Science** (M.Sc.) to **School of Data Science and Forecasting** affiliated to **Devi Ahilya Vishwavidyalaya University, Indore** is a record of the candidate own work carried out by her under **Dr. Vandit Hedau**. The matter embodied in this report is original and has not been submitted for the award of any other degree.

INTERNAL EXAMINER **EXTERNAL EXAMINER**

ACKNOWLEDGEMENT

I would like to express my heartfelt gratitude to all those who have supported me throughout the course of this project.

Firstly, I extend my sincere thanks to my mentors and colleagues for their valuable guidance, continuous encouragement, and insightful feedback, which played a critical role in shaping the outcomes of this project.

I am deeply grateful to my family and friends for their unwavering support and motivation, which provided me with the strength to consistently perform to the best of my abilities.

Lastly, I express my gratitude to my professors at the School of Data Science & Forecasting, DAVV. Special mentions to **Prof. Sanjiv Tokekar** (Head) and **Dr. Vandit Hedau** (Associate Professor), whose foundational teachings enabled me to fully leverage this internship opportunity. Their academic mentorship has been instrumental in developing my skills and nurturing my passion for data science and analytics.

This project marks a significant milestone in my academic and professional growth, and I sincerely thank everyone who contributed to its success.

NITISHAW SAINI

DS5B-2322

DECLARATION

This is to certify that Thesis entitled "**Sustainalyze: Automated ESG Risk Analyzer with AI Insights**" which is submitted by **NITISHAW SAINI** in partial fulfillment of the requirement for the award of degree of **Master in Science** (M.Sc.) **to School of Data Science & Forecasting** affiliated to **Devi Ahilya University**, **Indore** comprises only my own work and due acknowledgement has been made in the text to all other material used.

28th April 2025 **NITISHAW SAINI**

Table of Contents

1. Abstract

Environmental, Social, and Governance (ESG) considerations have become central to evaluating the long-term value and sustainability of companies. As the global focus shifts toward responsible investing and corporate accountability, there is an increasing demand for tools that can assess ESG risks in a structured, automated, and data-driven manner. Traditional financial models often overlook non-financial risks such as environmental impact, social controversies, or weak governance, which can significantly influence a company's reputation and long-term performance. Recognizing this gap, the *ESG Risk Analyzer* project was conceptualized to build a robust solution that combines data science, machine learning, and natural language processing (NLP) to analyse and predict ESG risks at scale.

The ESG Risk Analyzer collects and processes ESG-related data from companies listed on major indices such as the S&P Global 500 and Nifty 50. The workflow begins with data collection and cleaning, followed by feature engineering techniques that convert raw ESG metrics into meaningful risk indicators. A Random Forest Classifier model is trained to predict ESG risk labels based on quantitative inputs like ESG scores, risk exposure, controversy levels, and sector/industry encodings. For additional benchmarking, an XGBoost model is also implemented to compare performance.

To enrich the insights, the project integrates NLP methods such as sentiment analysis using TextBlob and controversy keyword extraction using tools like RAKE and Spacy. These helps identify key ESG concerns from company descriptions and public disclosures. Furthermore, GPT-style text generation techniques are used to produce concise, AI-generated ESG summaries that describe a company's overall ESG standing in human-readable format.

All components are integrated into an interactive dashboard built with Streamlit, allowing users to select a company, visualize ESG scores and trends, view machine learning predictions, and read AI-generated ESG insights. The dashboard provides clear risk exposure ratings, trend visualizations, and even a sample ESG report that can assist investors, analysts, and ESG officers in making informed decisions.

Overall, the ESG Risk Analyzer project demonstrates how data science, machine learning and NLP can be harnessed to enhance ESG risk assessment and improve transparency in ESG reporting. It provides a scalable framework that can be further extended with real-time data sources, forecasting models, and sector-specific risk analyses, making it a valuable contribution to the growing field of sustainable finance and responsible investing.

2. Introduction

2.1 Objective

The primary objective of the *ESG Risk Analyzer* project is to build an intelligent, end-to-end system that automates the analysis and visualization of Environmental, Social, and Governance (ESG) risks associated with publicly listed companies. ESG has become a critical dimension in investment and corporate evaluation, yet accessing meaningful ESG intelligence from raw datasets remains a complex task. This project aims to fill that gap by delivering a solution that transforms static ESG data into dynamic, interpretable insights.

By leveraging machine learning algorithms and natural language processing (NLP) techniques, the system categorizes companies based on ESG risk levels and identifies key risk indicators such as controversies or governance gaps. Additionally, it generates human-readable summaries using AI to help stakeholders interpret the results quickly and clearly.

The platform provides an interactive dashboard that allows users to visualize ESG risk scores, explore AI-generated insights, and analyze sector-wise benchmarks. The objective is not only to enhance transparency but also to empower stakeholders—such as investors, ESG consultants, financial analysts, and policymakers—with an effective tool for assessing sustainability risk and making data-driven decisions. Ultimately, the project envisions a shift from traditional ESG scorecards toward smarter, real-time ESG intelligence systems.

2.2 Purpose

The purpose of the *ESG Risk Analyzer* project is to transform raw, fragmented ESG data into meaningful, actionable insights that support ethical and sustainable investment decisions. While ESG datasets are increasingly being published, they often suffer from inconsistency, lack of structure, and limited interpretability. This creates challenges for stakeholders who need to evaluate non-financial risks across different companies and sectors.

This project aims to overcome these challenges by designing a system that automates data processing, risk classification, and insight generation using advanced techniques like machine learning and natural language processing (NLP). It processes structured ESG scores and unstructured textual information to generate comprehensive risk labels, extract key controversy keywords, and identify sentiment patterns. This multifaceted approach ensures that the final output is not only data-rich but also contextually relevant.

The purpose is also to democratize access to ESG insights through a user-friendly dashboard, enabling stakeholders to explore trends, benchmark performance, and interpret ESG risks interactively. By making ESG intelligence both accessible and insightful, the project contributes to better corporate accountability, informed investing, and responsible governance practices. In essence, the ESG Risk Analyzer seeks to elevate the role of ESG data in real-world decision-making.

2.3 Scope of the Project

The *ESG Risk Analyzer* is designed to serve as a comprehensive analytical framework that transforms raw ESG data into actionable intelligence. The scope of this project includes:

- **Analysing ESG risks** across companies based on structured ESG scores and unstructured textual data using machine learning and natural language processing techniques.
- **Integrating multi-source ESG datasets**, including corporate sustainability disclosures and sector-based benchmarks, to create a unified, enriched dataset.
- **Modelling ESG risk levels** using supervised learning algorithms like Random Forest and XGBoost to provide accurate, explainable predictions.
- **Extracting ESG-related keywords and sentiment** from company descriptions to uncover latent controversy signals and social sentiment patterns.
- **Generating personalized ESG insights** through GPT-style AI models that summarize a company's ESG posture in a concise and interpretable format.
- **Delivering interpretive dashboards** through Streamlit and Power BI to provide visual exploration tools, trend analysis, and comparative benchmarking across companies and industries.
- **Supporting decision-making** for investors, analysts, and ESG compliance officers by offering real-time predictions, risk indicators, and narrative insights in an intuitive interface.

This project is scoped to emphasize automation, interpretability, and accessibility of ESG risk insights, while laying the groundwork for future real-time data integration, sector-specific scoring models, and ESG report generation.

3. What is ESG?

ESG stands for Environmental, Social, and Governance—three critical pillars used to evaluate a company's ethical impact and sustainability practices. ESG metrics go beyond traditional financial performance indicators and assess how a company manages risks and opportunities related to environmental stewardship, social responsibility, and corporate governance. These factors are increasingly used by investors, analysts, and regulators to evaluate a company's long-term resilience, societal impact, and risk exposure.

Adopting ESG principles helps organizations align with sustainability goals, improve stakeholder trust, and build competitive advantage. In this context, ESG analysis has evolved into a key element of modern investment strategies and risk assessments, prompting the need for automated and intelligent ESG evaluation systems.

3.1 Environmental Factors

Environmental factors assess how a company interacts with and impacts the natural world. This includes its efforts to mitigate environmental risks and its commitment to sustainable practices. Key considerations under this pillar include:

- Carbon footprint and greenhouse gas emissions
- Energy efficiency and renewable energy usage
- Waste management and pollution control
- Water usage and conservation
- Biodiversity protection and deforestation policies

Companies with strong environmental performance often implement green technologies, reduce emissions, and adopt policies that promote sustainability throughout their operations and supply chains.

3.2 Social Factors

Social factors evaluate how a company manages relationships with employees, customers, suppliers, and the communities in which it operates. This pillar emphasizes human capital management, diversity, ethical practices, and social responsibility. Key metrics include:

- Labor standards and employee welfare
- Health and safety practices
- Diversity, equity, and inclusion (DEI)
- Customer satisfaction and data privacy
- Community engagement and philanthropy

Strong social performance indicates that a company prioritizes people and fosters a culture of respect, fairness, and ethical conduct, which can enhance reputation and stakeholder loyalty.

3.3 Governance Factors

Governance factors focus on the structures, policies, and practices that determine how a company is directed and controlled. This includes the quality of leadership, board independence, transparency, and shareholder rights. Common governance indicators are:

- Board structure and diversity
- Executive compensation alignment
- Anti-corruption and ethical practices
- Audit quality and internal controls
- Shareholder rights and disclosure standards

Effective governance ensures accountability, reduces the risk of unethical behavior, and aligns corporate practices with stakeholder interests, making it essential for sustainable long-term performance.

In conclusion, ESG—encompassing Environmental, Social, and Governance factors—is fundamental to assessing a company's sustainability and long-term impact. Sustainability, in this context, refers to the ability of businesses to operate responsibly while preserving natural resources, promoting social equity, and ensuring transparent governance. ESG analysis enables stakeholders to understand non-financial risks that could affect a company's reputation, compliance, and future growth. By integrating ESG principles into decision-making, companies not only contribute to a more sustainable world but also gain competitive advantage. As ESG continues to shape the future of investing and corporate accountability, it stands as a pillar of responsible and resilient growth.

4. Project Overview

4.1 Motivation

The global shift toward responsible investing has emphasized the importance of understanding a company's environmental, social, and governance practices. However, despite the abundance of ESG data, organizations often struggle to extract relevant insights due to inconsistent formats, lack of automation, and limited interpretability. This project was motivated by the need to bridge that gap—by developing a solution that processes both structured and unstructured ESG data, evaluates risk levels, and presents findings in a user-friendly, interactive format. The motivation stems from the vision of making ESG analysis more accessible, intelligent, and impactful for analysts, investors, and regulatory bodies.

4.2 Problem Statement

While ESG metrics are increasingly available, their utility is often limited due to challenges such as data inconsistency, lack of harmonization across sources, and minimal analytical depth. Most existing ESG platforms offer only static scores without context or explanation, leaving stakeholders with more questions than answers. Additionally, the absence of automation in extracting insights from ESG narratives and controversy reports creates barriers to timely decision-making. Thus, the core problem this project addresses is the lack of a scalable, interpretable, and real-time ESG risk analysis system that combines machine learning, NLP, and interactive visualizations.

4.3 Proposed Solution

The proposed solution is an end-to-end ESG Risk Analyzer that integrates data preprocessing, risk classification, and AI-powered insight generation into a single streamlined pipeline. It includes:

- Collection and harmonization of ESG data from S&P Global500 and Nifty 50 companies.
- Feature engineering to quantify ESG metrics, controversies, and sectoral performance.
- Supervised ML models (Random Forest and XGBoost) to classify ESG risk levels.
- NLP techniques like sentiment analysis, keyword extraction, and GPT-based text generation.
- Deployment of an interactive Streamlit dashboard and Power BI visualizations for intuitive ESG reporting.

This solution enables dynamic, accurate, and interpretable ESG insights, driving better sustainability evaluations and data-driven strategies.

5. Tech Stack Used

5.1 Backend

The backend of the ESG Risk Analyzer is developed in Python, providing a robust framework for data preprocessing, feature engineering, machine learning modeling, and natural language processing (NLP). Python's extensive libraries make it an ideal choice for handling complex ESG data and building predictive analytics systems. Data operations are managed using Pandas and NumPy, while machine learning tasks leverage Scikit-learn and XGBoost. For AI-driven ESG insights, a fine-tuned DistilGPT-2 model is employed to generate human-readable summaries based on company data and extracted keywords.

5.2 Frontend

The frontend is designed using **Streamlit**, enabling fast development of an interactive, web-based user interface. Through Streamlit, users can easily select companies, view ESG scores, risk classifications, controversy keywords, and AI-generated insights in real-time. In addition, a separate **Power BI dashboard** has been created to present detailed visualizations like sectoral benchmarks, ESG trends, and KPI-driven summaries, ensuring that both technical and business users can explore ESG data intuitively.

5.3 Libraries & Tools

The project utilizes a wide range of specialized libraries and tools:

- **Pandas & NumPy**: For efficient data manipulation and numerical operations.
- **Scikit-learn & XGBoost**: For building, training, and evaluating classification models.
- **Matplotlib, Seaborn & Plotly**: For generating static and interactive data visualizations.
- **TextBlob & spaCy**: For basic NLP tasks such as sentiment analysis and text preprocessing.
- **RAKE**: For extracting controversy-related keywords from company descriptions.
- **Transformers (Hugging Face)**: For AI insight generation using the **DistilGPT-2** language model, fine-tuned for ESG-specific summaries.
- **Joblib**: For saving and loading machine learning models.
- **Streamlit**: For building the interactive web dashboard application.
- **Power BI**: For advanced data visualization and dashboard reporting to support strategic ESG decision-making.

These technologies work together seamlessly to create an intelligent, end-to-end ESG analysis and reporting system.

6. Project Workflow

6.1 Data Collection

The foundation of the ESG Risk Analyzer project lies in the collection of high-quality, relevant ESG data. Data was sourced from reliable and authoritative platforms covering companies listed in the S&P Global 500 and Nifty 50 indices. The collected datasets included critical information such as Total ESG Risk Scores, Environment, Social, and Governance component scores, company descriptions, controversy scores, and sectoral classifications. In addition to real-world data, synthetic datasets were generated using the Faker library to simulate additional companies, thereby increasing the diversity and size of the training data without introducing bias. The goal of this phase was to ensure that the dataset captured a wide range of industries, ESG performance levels, and controversy types to build a resilient and generalizable machine learning model. Special care was taken to ensure the credibility and consistency of data sources, as ESG data can often be fragmented and vary across providers. The data collection step laid the groundwork for all downstream tasks, from machine learning to visualization, ensuring that the system had robust inputs for producing meaningful insights.

Extracting Data

[2]:
```
import pandas as pd
import numpy as np
```

[4]:
```
# --- Load datasets ---
sp500_df = pd.read_csv("SP 500 ESG Risk Ratings.csv")
nifty50_df = pd.read_csv("Nifty50.csv")
```

[6]:
```
# --- Standardize Column Names ---
sp500_df = sp500_df.rename(columns={
    "Name": "Company Name",
    "Total ESG Risk score": "Total_ESG_Risk_Score",
    "Environment Risk Score": "Environment_Score",
    "Governance Risk Score": "Governance_Score",
    "Social Risk Score": "Social_Score",
    "Controversy Level": "Controversy_Level",
    "Controversy Score": "Controversy_Score",
    "ESG Risk Percentile": "ESG_Risk_Percentile",
    "ESG Risk Level": "ESG_Risk_Level"
})

nifty50_df = nifty50_df.rename(columns={
    "company": "Company Name",
    "esg_risk_score_2024": "Total_ESG_Risk_Score",
    "predicted_future_esg_score": "Predicted_ESG_Score",
    "esg_risk_exposure": "ESG_Risk_Exposure",
    "esg_risk_management": "ESG_Risk_Management",
    "esg_risk_level": "ESG_Risk_Level",
    "Controversy Level": "Controversy_Level",
    "controversy_score": "Controversy_Score"
})
```

6.2 Data Cleaning & Preprocessing

Once the data was collected, it required extensive cleaning and preprocessing to be suitable for analysis and model training. Real-world datasets often contain missing values, duplicate records, inconsistent data formats, and errors, all of which can significantly affect model performance. For numerical columns like ESG scores, missing values were either imputed using domain knowledge or removed carefully to avoid bias. Categorical columns such as

Sector, Industry, and Controversy Level were encoded using Label Encoding to transform them into machine-readable numeric formats. Text data, particularly the company descriptions, underwent NLP preprocessing steps like lowercasing, removal of special characters, and tokenization to prepare them for sentiment analysis and keyword extraction. Special attention was given to harmonize fields across the different datasets (real and synthetic) so that feature columns matched perfectly. Additionally, outlier detection techniques were applied to ESG scores to ensure no extreme values skewed model learning. This preprocessing phase ensured that the final dataset was clean, structured, consistent, and ready for feature engineering and modeling tasks.

[12]:
```
# --- Generate values for missing columns ---
sp500_df = generate_missing_values(sp500_df)
nifty50_df = generate_missing_values(nifty50_df)
```

[14]:
```
# --- Fill any missing columns to match final schema ---
for col in final_columns:
    if col not in sp500_df.columns:
        sp500_df[col] = np.nan
    if col not in nifty50_df.columns:
        nifty50_df[col] = np.nan
```

[16]:
```
# --- Subset and reorder columns ---
sp500_clean = sp500_df[final_columns]
nifty50_clean = nifty50_df[final_columns]
```

[18]:
```
# --- Combine datasets ---
combined_esg_df = pd.concat([sp500_clean, nifty50_clean], ignore_index=True)
```

[20]:
```
# --- Save cleaned & merged dataset ---
combined_esg_df.to_csv("combined_esg_real_data.csv", index=False)
print("✅ Combined ESG dataset saved as 'combined_esg_real_data.csv'")
print("Combined shape:", combined_esg_df.shape)
```

```
✅ Combined ESG dataset saved as 'combined_esg_real_data.csv'
Combined shape: (553, 15)
```

Generating Synthetic Data

[23]:
```
from faker import Faker
import random

fake = Faker()
num_synthetic = 500

# --- Column Template ---
final_columns = [
    'Symbol', 'Company Name', 'Sector', 'Industry', 'Description',
    'Total_ESG_Risk_Score', 'Predicted_ESG_Score',
    'ESG_Risk_Exposure', 'ESG_Risk_Management', 'ESG_Risk_Level',
    'Environment_Score', 'Governance_Score', 'Social_Score',
    'Controversy_Level', 'Controversy_Score'
]

sectors = ["Technology", "Finance", "Healthcare", "Energy", "Consumer Goods", "Industrials"]
industries = ["Software", "Banks", "Pharmaceuticals", "Oil & Gas", "Retail", "Machinery"]

# --- Generator Function ---
def generate_synthetic_row():
    total_score = round(np.random.uniform(10, 50), 2)
    predicted_score = round(total_score + np.random.normal(0, 3), 2)
    controversy_score = random.randint(1, 100)
    controversy_level = (
        "Low" if controversy_score <= 20 else
        "Medium" if controversy_score <= 60 else "High"
    )
    return {
        'Symbol': fake.lexify(text='???'),
        'Company Name': fake.company(),
        'Sector': random.choice(sectors),
        'Industry': random.choice(industries),
        'Description': fake.catch_phrase(),
        'Total_ESG_Risk_Score': total_score,
```

```
# --- Combine with real dataset ---
combined_all_df = pd.concat([combined_esg_df, synthetic_df], ignore_index=True)

# --- Save final dataset ---
combined_all_df.to_csv("final_esg_dataset.csv", index=False)
print("✅ Final ESG dataset with synthetic data saved as 'final_esg_dataset.csv'")
print("Final shape:", combined_all_df.shape)
```

```
✅ Final ESG dataset with synthetic data saved as 'final_esg_dataset.csv'
Final shape: (1053, 15)
```

6.3 Feature Engineering

Feature engineering was a critical phase where raw data was transformed into meaningful inputs that enhanced the predictive power of machine learning models. New features such as Total ESG Risk Score were created by aggregating individual Environment, Social, and Governance scores. Companies were categorized into ESG Risk Levels (Low, Moderate, High) based on Total ESG Risk Score thresholds, creating the target variable for the classification models. In addition to structured features, sentiment scores were generated from company descriptions using TextBlob, adding a text-based dimension to the structured dataset. Categorical fields like Sector, Industry, and Controversy Level were numerically encoded to facilitate learning by tree-based models like Random Forest and XGBoost. Important insights such as controversy scores and their corresponding levels were treated as strong predictors, considering that controversies heavily influence a company's ESG risk profile. The feature engineering phase also included standardizing feature ranges and handling class imbalance, if any, to ensure fair learning across different ESG risk categories. Well-engineered features contributed significantly to improving model accuracy and interpretability, ultimately making the predictions more reliable.

Feature Engineering For ML Model

[ ]:
```
import pandas as pd
import numpy as np

# --- Load combined dataset ---
df = pd.read_csv("final_esg_dataset.csv")
```

[2]:
```
df.columns
```

[2]:
```
Index(['Symbol', 'Company Name', 'Sector', 'Industry', 'Description',
       'Total_ESG_Risk_Score', 'Predicted_ESG_Score', 'ESG_Risk_Exposure',
       'ESG_Risk_Management', 'ESG_Risk_Level', 'Environment_Score',
       'Governance_Score', 'Social_Score', 'Controversy_Level',
       'Controversy_Score'],
      dtype='object')
```

[3]:
```
# --- Clean numeric columns ---
numeric_cols = [
    'Total_ESG_Risk_Score', 'Predicted_ESG_Score', 'ESG_Risk_Exposure',
    'ESG_Risk_Management', 'Environment_Score', 'Governance_Score',
    'Social_Score', 'Controversy_Score'
]

for col in numeric_cols:
    df[col] = pd.to_numeric(df[col], errors='coerce')
```

[7]:
```
# --- Drop rows with invalid numeric values (optional, safe step) ---
df = df.dropna(subset=numeric_cols).reset_index(drop=True)
```

[9]:
```python
# --- ESG Risk Labeling ---
def label_risk(score):
    if score <= 20:
        return "Low"
    elif score <= 40:
        return "Medium"
    else:
        return "High"

df['ESG_Risk_Label'] = df['Total_ESG_Risk_Score'].apply(label_risk)
```

[11]:
```python
# --- Confirm label distribution ---
print("Label counts:")
print(df['ESG_Risk_Label'].value_counts())
```

```
Label counts:
ESG_Risk_Label
Medium    475
Low       281
High      147
Name: count, dtype: int64
```

[13]:
```python
# --- Save feature-engineered dataset ---
df.to_csv("final_esg_dataset_labeled.csv", index=False)
print("✅ Feature-engineered dataset saved as 'final_esg_dataset_labeled.csv'")
print("Final shape:", df.shape)
```

```
✅ Feature-engineered dataset saved as 'final_esg_dataset_labeled.csv'
Final shape: (903, 16)
```

6.4 Machine Learning Modelling

After feature engineering, the dataset was used to train machine learning models to predict ESG risk categories. The primary model selected was the Random Forest Classifier, known for its ability to handle structured data efficiently, resist overfitting, and provide interpretable feature importance scores. Hyperparameters like the number of trees and maximum depth were tuned to optimize model performance. XGBoost, a more advanced boosting-based algorithm, was also implemented for comparison to evaluate whether it could outperform Random Forest in terms of precision, recall, and F1-score. Models were evaluated using stratified train-test splits to ensure each ESG Risk Level was represented proportionately. Metrics such as Accuracy, Precision, Recall, F1-Score, Mean Absolute Error (MAE), and Mean Squared Error (MSE) were used for a holistic evaluation. Feature importance analysis further confirmed the validity of selected features like Controversy Score, Sector, and Governance Score. The best-performing model (Random Forest) was serialized using Joblib, allowing it to be seamlessly deployed in the Streamlit application for real-time ESG risk prediction. Machine learning modeling turned the static ESG data into actionable predictions, providing critical support for decision-makers.

ML Model training

```
[ ]: import pandas as pd
     import numpy as np
     from sklearn.model_selection import train_test_split
     from sklearn.ensemble import RandomForestClassifier
     from xgboost import XGBClassifier
     from sklearn.metrics import classification_report, accuracy_score, confusion_matrix
```

```
[2]: # --- Load feature-engineered dataset ---
     df = pd.read_csv("final_esg_dataset_labeled.csv")
```

```
[3]: # --- Define features and target ---
     features = [
         'Total_ESG_Risk_Score', 'Predicted_ESG_Score', 'ESG_Risk_Exposure',
         'ESG_Risk_Management', 'Environment_Score', 'Governance_Score',
         'Social_Score', 'Controversy_Score'
     ]
     target = 'ESG_Risk_Label'

     X = df[features]
     y = df[target]
```

```
[4]: # --- Encode labels ---
     from sklearn.preprocessing import LabelEncoder
     le = LabelEncoder()
     y_encoded = le.fit_transform(y)
```

```
[5]: # Save for decoding later
     label_map = dict(zip(le.classes_, le.transform(le.classes_)))
     print("Label Encoding:", label_map)
```

```
Label Encoding: {'High': 0, 'Low': 1, 'Medium': 2}
```

```
[6]: # --- Split data ---
     X_train, X_test, y_train, y_test = train_test_split(X, y_encoded, test_size=0.2, random_state=42, stratify=y_encoded)
```

```
[7]: # --- Train Random Forest ---
     rf = RandomForestClassifier(n_estimators=100, random_state=42)
     rf.fit(X_train, y_train)
     rf_preds = rf.predict(X_test)
```

```
[8]: # --- Train XGBoost ---
     xgb = XGBClassifier(use_label_encoder=False, eval_metric='mlogloss', random_state=42)
     xgb.fit(X_train, y_train)
     xgb_preds = xgb.predict(X_test)
```

```
[9]: # --- Evaluation ---
     print("🔍 Random Forest Results:")
     print(classification_report(y_test, rf_preds, target_names=le.classes_))
     print("Confusion Matrix:\n", confusion_matrix(y_test, rf_preds))

     print("\n🔍 XGBoost Results:")
     print(classification_report(y_test, xgb_preds, target_names=le.classes_))
     print("Confusion Matrix:\n", confusion_matrix(y_test, xgb_preds))
```

```
:  # --- Save the best model
   import joblib
   joblib.dump(rf, "rf_esg_model.pkl")
   joblib.dump(le, "label_encoder.pkl")
   print("✅ Saved Random Forest model and label encoder for Streamlit app.")
```

```
✅ Saved Random Forest model and label encoder for Streamlit app.
```

6.5 NLP & AI Insight Generation

Natural Language Processing (NLP) played a pivotal role in generating qualitative insights from company descriptions. The first step was controversy keyword extraction using RAKE which identified the top controversy-related phrases from the textual data. These keywords provided a glimpse into the specific environmental, social, or governance issues faced by the company. Sentiment analysis using TextBlob determined the overall tone of the company's public perception, adding another layer of context. To make the insights accessible and easily

interpretable, a DistilGPT-2 model (a lightweight version of GPT-2) was fine-tuned on ESG-specific narratives. The model took structured data and extracted keywords as input to generate two-line ESG summaries for each company. These AI-generated insights provided users with a quick but meaningful understanding of a company's ESG profile without having to sift through raw data or lengthy reports. This NLP and AI layer transformed otherwise complex ESG information into intuitive, human-readable summaries, significantly enhancing the project's usability and impact.

AI insight generation with distilgpt2

[45]:
```python
import pandas as pd
from transformers import AutoTokenizer, AutoModelForCausalLM
from tqdm import tqdm

# Load data
df = pd.read_csv("final_esg_dataset_with_structured_insights.csv")
df["Description"] = df["Description"].fillna("No description available.")

# Load tokenizer & model
tokenizer = AutoTokenizer.from_pretrained("distilgpt2")
model = AutoModelForCausalLM.from_pretrained("distilgpt2")

# Set padding token for GPT-2
tokenizer.pad_token = tokenizer.eos_token
model.config.pad_token_id = tokenizer.eos_token_id

# Insight generation function
def generate_esg_insight(description, risk_level):
    prompt = f"ESG Risk Level: {risk_level}. Description: {description}. Insight:"
    inputs = tokenizer(prompt, return_tensors="pt", truncation=True, padding=True)

    outputs = model.generate(
        inputs["input_ids"],
        attention_mask=inputs["attention_mask"],
        max_new_tokens=40,
        do_sample=True,
        top_k=50,
        top_p=0.95,
        temperature=0.7,
        pad_token_id=tokenizer.eos_token_id
    )

    generated_text = tokenizer.decode(outputs[0], skip_special_tokens=True)
    return generated_text.split("Insight:")[-1].strip() if "Insight:" in generated_text else generated_text.strip()
```

```python
    generated_text = tokenizer.decode(outputs[0], skip_special_tokens=True)
    return generated_text.split("Insight:")[-1].strip() if "Insight:" in generated_text else generated_text.strip()

# Add ESG_AI_Insight column using progress bar
tqdm.pandas(desc="🔍 Generating ESG AI Insights")
df["ESG_AI_Insight"] = df.progress_apply(lambda row: generate_esg_insight(row["Description"], row["ESG_Risk_Level"]), axis=1)
```

```
🔍 Generating ESG AI Insights: 100%|██████████| 903/903 [29:53<00:00,  1.99s/it]
```

[48]:
```python
# Save to CSV
df.to_csv("final_esg_dataset_with_gpt2_insights.csv", index=False)
print("✅ ESG insights generated and saved with progress bar.")
```

```
✅ ESG insights generated and saved with progress bar.
```

Sentiment Analysis of Description and Controversy

[62]:
```
from textblob import TextBlob

df['Description_Sentiment'] = df['Description'].apply(lambda x: TextBlob(x).sentiment.polarity)
df['Description_Sentiment_Label'] = df['Description_Sentiment'].apply(
    lambda x: 'Positive' if x > 0.1 else ('Negative' if x < -0.1 else 'Neutral')
)
```

[63]:
```
#controversy
# If you've already added controversy terms, skip this. Otherwise, ensure it's included.
# Example fallback if not extracted yet:
df['Top_Controversy_Terms'] = df['Description'].apply(lambda x: ["no major controversies found"] if pd.isna(x) else [x])

# --- 🧠 NLP Insight Generator Function (Template-Based) ---
def generate_esg_insight(row):
    risk = row['ESG_Risk_Level']
    terms = row['Top_Controversy_Terms']

    # If string, try splitting it into a list
    if isinstance(terms, str):
        terms = [term.strip() for term in terms.split(",")]

    # Build short explanation
    if risk == 'High':
        return f"The company faces high ESG risk due to major concerns like {', '.join(terms[:2])}."
    elif risk == 'Medium':
        return f"The company faces moderate ESG risk driven by issues such as {', '.join(terms[:2])}."
    else:
        return f"The company shows low ESG risk with minor concerns like {', '.join(terms[:2])}."

# --- 🧠 Apply the Insight Generator to the DataFrame ---
df['ESG_AI_Insight'] = df.apply(generate_esg_insight, axis=1)
```

[66]:
```
# Save the new dataset with insights
df.to_csv("final_esg_dataset_with_structured_insights.csv", index=False)
print("✅ Structured AI ESG insights saved successfully!")
```

```
✅ Structured AI ESG insights saved successfully!
```

[164]:
```
pd.read_csv("final_esg_dataset_with_structured_insights.csv").head(2)
```

[164]:

	Symbol	Company Name	Sector	Industry	Description	Total_ESG_Risk_Score	Predicted_ESG_Score	ESG_Risk_Exposure	ESG_Risk_Management	ESG_Risk_Level	...	Co
0	EMN	Eastman Chemical Company	Basic Materials	Specialty Chemicals	Eastman Chemical Company operates as a special...	25.3	15.548391	32.795002	42.033837	Medium	...	
1	DPZ	Domino's Pizza Inc.	Consumer Cyclical	Restaurants	Domino's Pizza, Inc., through its subsidiaries...	29.2	23.613286	62.572422	33.762561	Medium	...	

2 rows × 24 columns

RAKE

[166]:
```
#!pip install rake-nltk
```

[120]:
```
from rake_nltk import Rake
from textblob import TextBlob
import pandas as pd

# Initialize RAKE
r = Rake()

# Clean descriptions
df['Description'] = df['Description'].apply(lambda x: x if isinstance(x, str) and x.strip() != '' else 'No ESG description available')

# Sentiment Analysis
df['Description_Sentiment'] = df['Description'].apply(lambda x: TextBlob(x).sentiment.polarity)
df['Description_Sentiment_Label'] = df['Description_Sentiment'].apply(
    lambda x: 'Positive' if x > 0.1 else ('Negative' if x < -0.1 else 'Neutral')
)

# Keyword Extraction using RAKE
def extract_keywords_rake(text):
    try:
        r.extract_keywords_from_text(text)
        keywords = r.get_ranked_phrases()[:3]
        return keywords if keywords else ['no major controversies found']
    except:
        return ['no major controversies found']

df['Top_Controversy_Terms'] = df['Description'].apply(extract_keywords_rake)

# AI ESG Insight Generation
def generate_esg_insight(row):
    risk = row.get('ESG_Risk_Level', 'Medium')
    terms = row.get('Top_Controversy_Terms', [])
    if not isinstance(terms, list):
        terms = [terms]
    top_terms = ', '.join(terms[:2])
    if risk in ['High', 'Severe']:
        return f"The company faces {risk.lower()} ESG risk driven by issues such as {top_terms}."
    else:
        return f"The company shows low ESG risk with minor concerns like {top_terms}."

df['ESG_AI_Insight'] = df.apply(generate_esg_insight, axis=1)
```

[122]:
```
df.head()  # Shows the top rows with all new columns

# OR preview only the most relevant columns
df[['Company Name', 'Description_Sentiment_Label', 'Top_Controversy_Terms', 'ESG_AI_Insight']].head(10)
```

[122]:

	Company Name	Description_Sentiment_Label	Top_Controversy_Terms	ESG_AI_Insight
0	Eastman Chemical Company	Positive	[functional products segment offers amine deri...	The company faces moderate ESG risk driven by ...
1	Domino's Pizza Inc.	Negative	[pepperoni stuffed cheesy breads, soft drink p...	The company faces moderate ESG risk driven by ...
2	Davita Inc.	Neutral	[company operates kidney dialysis centers, com...	The company faces moderate ESG risk driven by ...
3	Darden Restaurants, Inc.	Neutral	[capital burger brand names, inc ., together, ...	The company faces moderate ESG risk driven by ...
4	Zoetis Inc.	Neutral	[company commercializes products primarily acr...	The company shows low ESG risk with minor conc...
5	Zimmer Biomet Holdings, Inc.	Neutral	[thoracic products comprising face, medical te...	The company faces moderate ESG risk driven by ...
6	Yum! Brands, Inc.	Neutral	[franchises quick service restaurants worldwid...	The company faces moderate ESG risk driven by ...
7	Xylem Inc	Neutral	[control solutions segment offers smart meters...	The company shows low ESG risk with minor conc...
8	Xcel Energy, Inc.	Neutral	[xcel energy inc ., leases natural gas pipeli...	The company faces moderate ESG risk driven by ...
9	Wynn Resorts Ltd	Neutral	[encore boston harbor segment operates casino ...	The company faces moderate ESG risk driven by ...

[124]:
```
# Save the new dataset with insights
df.to_csv("final_esg_dataset_with_structured_insights_withsentiments.csv", index=False)
print("✅ saved successfully!")
```

```
✅ saved successfully!
```

USING SPACY

[150]:
```
#pip install spacy
#python -m spacy download en_core_web_sm
```

[152]:
```
import spacy
from tqdm import tqdm

# Load spaCy model
nlp = spacy.load("en_core_web_sm")

# Keyword extractor using noun chunks (phrases)
def extract_keywords_spacy(text):
    try:
        text = str(text).strip()
        if not text or text.lower() in ['nan', 'none', 'null']:
            return ['no major controversies found']

        doc = nlp(text)
        # Extract noun phrases as keywords
        keywords = list(set(chunk.text.strip() for chunk in doc.noun_chunks if len(chunk.text.strip()) > 2))
        return keywords[:3] if keywords else ['no major controversies found']

    except Exception as e:
        return [f'error: {str(e)}']
```

[154]:
```
tqdm.pandas(desc="🧠 Extracting with spaCy")
df['Top_Controversy_Terms'] = df['Description'].progress_apply(extract_keywords_spacy)
```

```
🧠 Extracting with spaCy: 100%|| 903/903 [00:26<00:00, 34.13it/s]
```

[174]:
```
df['ESG_AI_Insight_spacy'] = df.apply(generate_esg_insight_spacy, axis=1)
```

[176]:
```
# Save the new dataset with insights
df.to_csv("final_esg_dataset_with_spacy_insights.csv", index=False)
print("✅ saved successfully!")
```

```
✅ saved successfully!
```

[178]:
```
df.head()
```

[178]:

ry_encoded	Controversy_Level_encoded	AI_ESG_Insight	Description_Sentiment	Description_Sentiment_Label	Top_Controversy_Terms	ESG_AI_Insight	ESG_AI_Insight_spacy
99	5	The company operates in the Basic Materials se...	0.166667	Positive	[organic acid-based solutions, polyvinyl butyr...	The company faces moderate ESG risk driven by ...	The company faces moderate ESG risk driven by ...
90	5	The company operates in the Consumer Cyclical ...	-0.214286	Negative	[three segments, Ann Arbor, Michigan]	The company faces moderate ESG risk driven by ...	The company faces moderate ESG risk driven by ...
63	5	The company operates in the Healthcare sector ...	0.023333	Neutral	[acute inpatient dialysis services, kidney dia...	The company faces moderate ESG risk driven by ...	The company faces moderate ESG risk driven by ...
90	5	The company operates in the Consumer Cyclical ...	0.000000	Neutral	[Darden Restaurants, Eddie V's Prime Seafood, ...	The company faces moderate ESG risk driven by ...	The company faces moderate ESG risk driven by ...

6.6 Streamlit Dashboard Integration

The final phase of the workflow involved integrating the results into an interactive, user-friendly dashboard built using Streamlit. The dashboard allowed users to select a company from a dropdown list and view its ESG Risk Score, Predicted ESG Score, Risk Exposure, Risk Management, Controversy Score, Risk Label, Sector, Industry, extracted Controversy Keywords, and AI-Generated ESG Insight, all in a well-organized layout. Real-time model inference was incorporated by loading the trained Random Forest model to predict ESG risk levels dynamically based on user selection. In addition to the Streamlit dashboard, a detailed

Power BI dashboard was also created to offer advanced visual analytics. The Power BI dashboard includes ESG KPI cards, sector and industry benchmarks, ESG score trend charts, heatmaps of controversy incidents, and downloadable ESG reports. Together, Streamlit and Power BI deliver a full-stack experience, combining machine learning predictions, AI insights, and rich visualizations to empower ESG analysts, investors, and stakeholders with actionable intelligence.

```python
import streamlit as st
import pandas as pd
import plotly.express as px
import joblib

# Load Data and Model
df = pd.read_csv("final_esg_dataset_with_spacy_insights.csv")
model = joblib.load("rf_esg_model.pkl")

# Sidebar - Company Selection
st.sidebar.title("📊 ESG Risk Analyzer")

# 🎯 Sidebar Filters
st.sidebar.header("🔎 Filter Companies")

# Filter widgets: Sector and Risk Level only
selected_sector = st.sidebar.selectbox("Select Sector", ['All'] + sorted(df['Sector'].dropna().unique().tolist()))
selected_risk = st.sidebar.selectbox("Select ESG Risk Level", ['All'] + sorted(df['ESG_Risk_Level'].dropna().unique().tolist()))

# Apply filters to create a filtered DataFrame
filtered_df = df.copy()
if selected_sector != 'All':
    filtered_df = filtered_df[filtered_df['Sector'] == selected_sector]
if selected_risk != 'All':
    filtered_df = filtered_df[filtered_df['ESG_Risk_Level'] == selected_risk]

# Display number of matching results
st.sidebar.markdown(f"🔢 **Filtered Results:** {len(filtered_df)} Companies")

# ✅ Check for empty filter result
if filtered_df.empty:
    st.warning(f"⚠️ No companies available for Sector **{selected_sector}** and ESG Risk Level **{selected_risk}**.")
    st.stop()  # 🔴 Stops further rendering
else:
    # Dynamic company dropdown based on filters
    company = st.sidebar.selectbox("Choose a Company", sorted(filtered_df['Company Name'].unique()))
```

```python
    # Pull selected company data
    info = filtered_df[filtered_df["Company Name"] == company].iloc[0]
selected_row = df[df['Company Name'] == company].iloc[0]


# Header
st.title("🌍 ESG Risk Dashboard")
st.markdown(f"### Company: **{company}**")



# Extract input features from the selected company's row
input_features = [
    info['Total_ESG_Risk_Score'],
    info['Predicted_ESG_Score'],
    info['ESG_Risk_Exposure'],
    info['ESG_Risk_Management'],
    info['Controversy_Score'],
    info['Sector_encoded'],
    info['Industry_encoded'],
    info['Controversy_Level_encoded']
]

predicted_label = model.predict([input_features])[0]

# Map label to risk level and emoji
risk_level_map = {
    0: ("Low", "🟢"),
    1: ("Medium", "🟠"),
    2: ("High", "🔴")
}
risk_label, color_emoji = risk_level_map.get(predicted_label)

# ✅ Update info dict so KPI uses this
info['ESG_Risk_Label'] = risk_label
```

.

.

.

.

```python
# Function to generate PDF
def generate_pdf_report(df_row):
    pdf = FPDF()
    pdf.add_page()
    pdf.set_font("Arial", size=12)

    pdf.set_text_color(0, 102, 204)
    pdf.set_font("Arial", 'B', 16)
    pdf.cell(200, 10, txt="ESG Risk Report", ln=True, align='C')
    pdf.ln(10)

    pdf.set_font("Arial", size=12)
    pdf.set_text_color(0, 0, 0)

    for col, val in df_row.items():
        pdf.multi_cell(0, 10, txt=f"{col}: {val}", border=0)

    return pdf

# ---14. Generate and Download PDF ESG Report ---
if st.button("📄 Generate ESG PDF Report"):
    selected_row = df[df['Company Name'] == company].iloc[0]
    pdf = generate_pdf_report(selected_row)

    # Save to a temporary file
    with tempfile.NamedTemporaryFile(delete=False, suffix=".pdf") as tmp_file:
        pdf.output(tmp_file.name)
        with open(tmp_file.name, "rb") as f:
            st.download_button(
                label="📥 Download ESG Report (PDF)",
                data=f,
                file_name=f"{company}_esg_report.pdf",
                mime="application/pdf"
            )

```

7. Key Techniques Used

7.1 Random Forest Classifier

The Random Forest Classifier served as the primary supervised learning model for categorizing companies into ESG risk levels. Random Forest is an ensemble method that constructs multiple decision trees during training and merges their outputs to improve classification accuracy and control overfitting. It was selected due to its robustness, ease of interpretability, and capability to handle complex feature interactions across environmental, social, and governance metrics. Hyperparameters such as the number of trees (n_estimators) and maximum depth (max_depth) were fine-tuned for optimal performance. The model demonstrated strong results in terms of accuracy, recall, and precision. Feature importance rankings from Random Forest helped validate the significance of critical indicators like Controversy Level, Governance Score, and Risk Management measures. The model was serialized and integrated into the Streamlit dashboard to perform real-time predictions, ensuring users could easily assess ESG risks with just a few clicks.

7.2 XGBoost Comparison

To ensure the reliability and competitiveness of the machine learning results, the XGBoost Classifier was also trained and evaluated. XGBoost (Extreme Gradient Boosting) is a highly efficient and scalable gradient boosting algorithm that often outperforms other models in structured data scenarios. During experimentation, XGBoost slightly edged out Random Forest on some performance metrics, offering high accuracy and better handling of class imbalances. However, Random Forest was ultimately chosen for deployment because of its faster inference speed and interpretability, which are essential in real-time dashboard applications. The XGBoost results served as a valuable benchmark, reinforcing confidence in the Random Forest model while demonstrating that alternative advanced techniques were also considered during model selection.

7.3 Sentiment Analysis (TextBlob)

Sentiment analysis was utilized to gauge public perception related to companies based on their textual descriptions. TextBlob, a straightforward and widely-used Python library, was employed for this task. It computes two key scores: polarity (ranging from -1 for very negative to +1 for very positive sentiment) and subjectivity (ranging from 0 for objective to 1 for highly subjective content). These scores provided additional features that enriched the ESG risk prediction pipeline. Positive sentiment generally aligned with strong governance, good environmental practices, or positive social initiatives, while negative sentiment often pointed toward scandals, controversies, or corporate irresponsibility. Including sentiment analysis added a qualitative layer to the quantitative ESG data, enabling a more comprehensive assessment of corporate behavior beyond raw metrics.

7.4 Controversy Keyword Extraction (RAKE / spaCy)

Extracting key phrases related to controversies was crucial for enhancing the understanding of ESG risks. A combination of RAKE (Rapid Automatic Keyword Extraction) and spaCy was

used for this purpose. RAKE quickly identified important multi-word keywords by analyzing the frequency and co-occurrence of words in the company descriptions. It provided a lightweight, unsupervised method for capturing essential controversy-related terms. Meanwhile, spaCy's NLP capabilities — such as tokenization, part-of-speech tagging, and stopword removal — ensured the text was clean and well-structured before keyword extraction. This two-step approach enabled the identification of core ESG issues (e.g., "oil spill," "labor violation," "governance failure") associated with each company. The extracted controversy terms were displayed in the dashboard, offering users immediate insights into the types of risks a company might be facing without having to read full-length descriptions.

7.5 GPT-style Insight Generation

To add meaningful, human-readable insights to the ESG Risk Analyzer, we implemented **AI-generated text summaries** using **DistilGPT-2**, a transformer model available through the Hugging Face library. DistilGPT-2 is a **distilled version of GPT-2**, making it **lighter, faster**, and **more efficient** while maintaining high-quality text generation capabilities. This makes it highly suitable for projects requiring scalable and responsive AI outputs.

In our workflow, after analyzing ESG-related metrics, controversies, and risk scores for each company, the processed features were fed into DistilGPT-2. The model then generated concise, interpretative sentences summarizing the company's ESG standing. These AI-generated insights help stakeholders quickly understand ESG risk exposure without needing to dig through raw numbers or lengthy reports.

Using DistilGPT-2 offers several benefits:

- **Speed and Efficiency**: Faster inference compared to the full GPT-2 model.
- **Lightweight Deployment**: Easier to integrate within applications like Streamlit dashboards without heavy computational overhead.
- **Accuracy**: Produces contextually relevant and coherent outputs suitable for ESG reporting.
- **Ease of Use**: Directly accessible via the Hugging Face transformers library with minimal setup.

Through this technique, the ESG Risk Analyzer was able to automatically create brief, meaningful ESG summaries for hundreds of companies, enhancing the dashboard's interpretability and user experience.

8. Results & Visualization

8.1 ESG Risk Score Visualization

To help stakeholders easily interpret ESG risk exposure, the project incorporates dynamic visualizations of ESG scores across companies. Using Matplotlib, Seaborn, and Plotly, various charts were created to show:

- Total ESG Risk Scores
- Individual Environmental, Social, and Governance Sub scores
- Controversy Scores and Levels

Bar charts and color-coded risk meters were used to make the visualization more intuitive. High-risk companies were shown in red hues, moderate-risk companies in orange, and low-risk companies in green, ensuring instant recognition of ESG standings. Additionally, trend graphs were implemented wherever historical ESG data was available, demonstrating score evolution over time. These visualizations enabled users to compare companies across sectors and industries effectively and spot potential ESG concerns quickly.

8.2 AI-Generated Insights

An innovative highlight of the ESG Risk Analyzer is the integration of AI-Generated ESG Insights. Using the DistilGPT-2 model, custom text summaries were generated based on each company's ESG data and controversy analysis. These insights condense complex ESG profiles into 1-2 meaningful sentences.

For instance:

"Company ABC exhibits strong ESG performance overall but faces moderate risks due to governance-related controversies."

These AI-driven summaries were dynamically displayed alongside the ESG score visualizations in the dashboard, offering both numerical and textual perspectives for enhanced decision-making.
The use of Natural Language Generation (NLG) here bridges the gap between raw data and human understanding, making the tool highly practical for investors and analysts.

8.3 Interactive Dashboard Screenshots

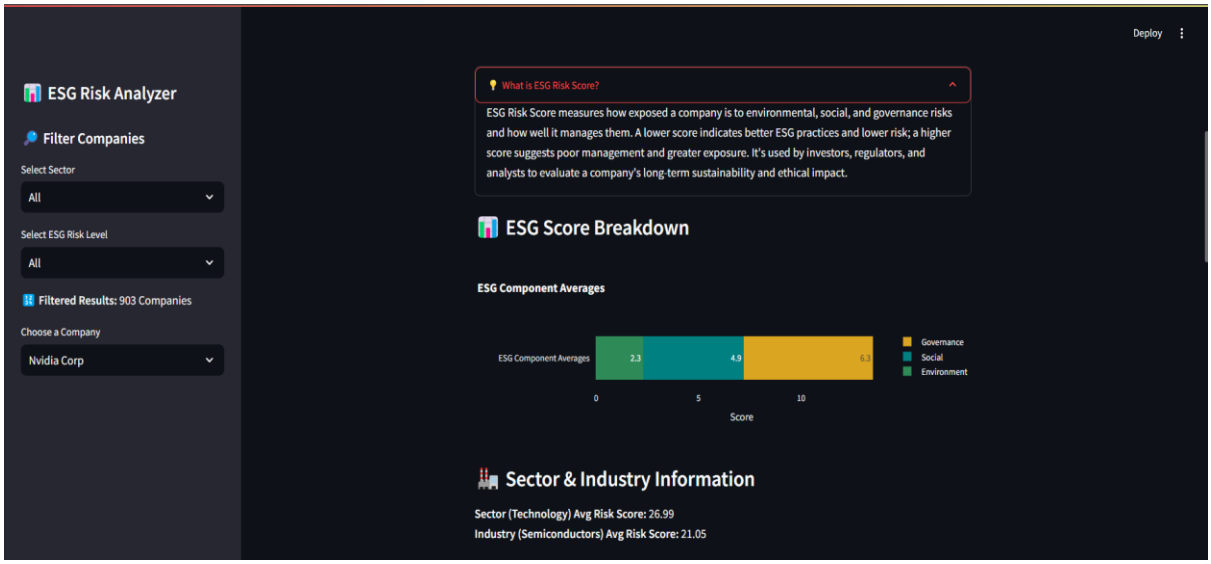
The final results were showcased through an interactive Streamlit and Power BI Dashboard:

- Streamlit Dashboard: Provided live prediction capabilities, AI insights, ESG KPIs, and visual score comparisons across companies.
- Power BI Dashboard: Enabled more detailed and customizable reporting with slicers, filters, KPI cards, sector/industry-wise ESG comparisons, and beautiful visual storytelling.

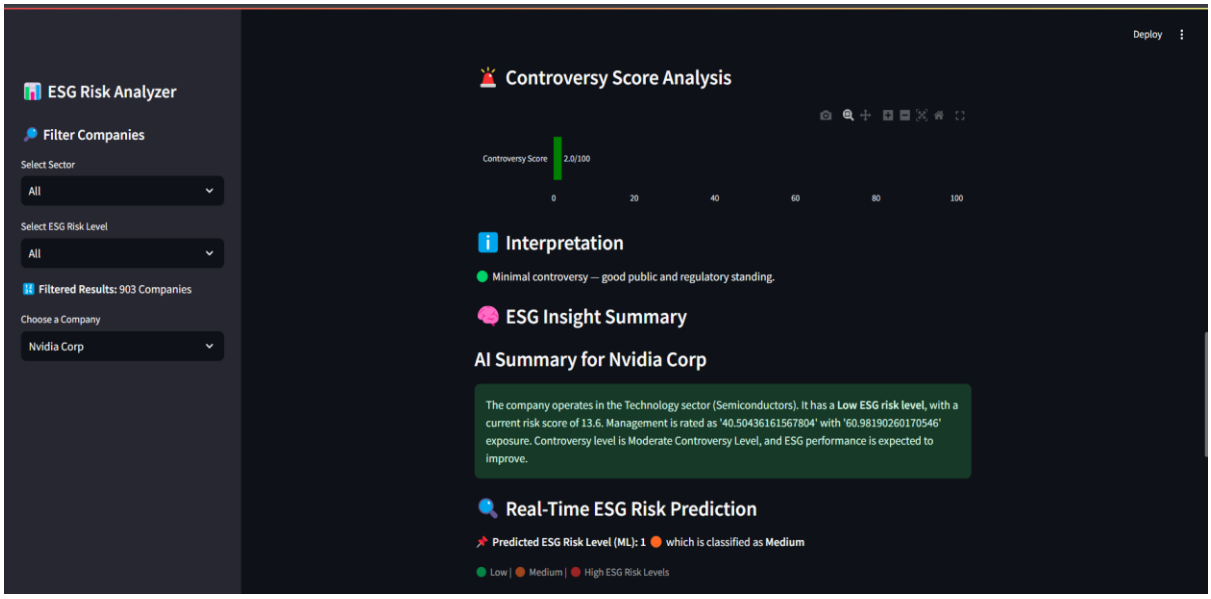
The dashboard screenshots include:

- ESG Score Overview Panels
- Risk Exposure Meters
- AI-Generated ESG Insight Sections
- Controversy Keyword Clouds
- Company Risk Level Comparisons (Grouped by Sector/Industry)

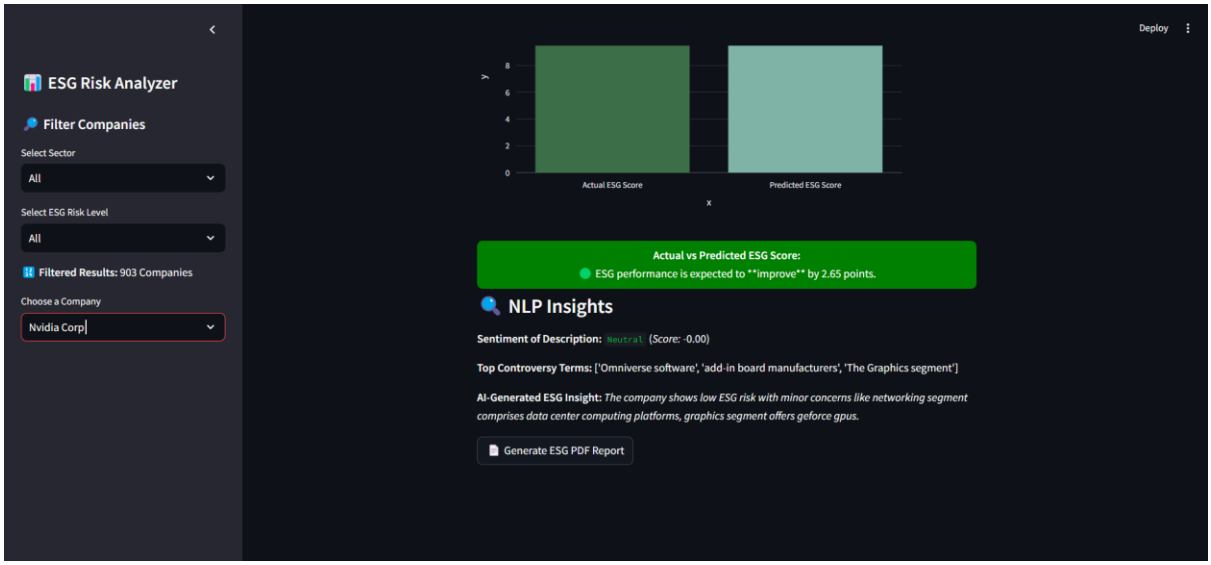
These dashboards make the ESG Risk Analyzer not only a machine learning project but also a full-fledged decision-support system ready for real-world application.




ESG Risk Analyzer
Filter Companies
Select Sector
All
Select ESG Risk Level
All
Filtered Results: 903 Companies
Choose a Company
Nvidia Corp
Deploy
ESG Score Comparison: Company vs Sector vs Industry
ESG Score Comparison for Nvidia Corp
ESG Score
Category
Selected Company
Technology Avg
Semiconductors Avg
13.6
26.98887
21.05455
Insight: Nvidia Corp demonstrates stronger ESG performance compared to its sector (Technology) and industry (Semiconductors). The ESG risk score is 13.4 points lower than the sector average and 7.5 points lower than the industry average.


ESG Risk Analyzer
Filter Companies
Select Sector
All
Select ESG Risk Level
All
Filtered Results: 903 Companies
Choose a Company
Nvidia Corp
Deploy
Controversy Score Analysis
Controversy Score
2.0/100
Interpretation
Minimal controversy — good public and regulatory standing.
ESG Insight Summary
AI Summary for Nvidia Corp
The company operates in the Technology sector (Semiconductors). It has a Low ESG risk level, with a current risk score of 13.6. Management is rated as '40.50436161567804' with '60.98190260170546' exposure. Controversy level is Moderate Controversy Level, and ESG performance is expected to improve.
Real-Time ESG Risk Prediction
Predicted ESG Risk Level (ML): 1 which is classified as Medium
Low | Medium | High ESG Risk Levels


ESG Risk Analyzer
Filter Companies
Select Sector
All
Select ESG Risk Level
All
Filtered Results: 903 Companies
Choose a Company
Nvidia Corp
Deploy
Actual vs Predicted ESG Score
Actual vs Predicted ESG Score for Nvidia Corp
color
Actual
Predicted
13.6
10.951
Actual ESG Score
Predicted ESG Score
Actual vs Predicted ESG Score:
ESG performance is expected to **improve** by 2.65 points.


ESG Risk Analyzer
Filter Companies
Select Sector
All
Select ESG Risk Level
All
Filtered Results: 903 Companies
Choose a Company
Nvidia Corp
Deploy
Actual ESG Score
Predicted ESG Score
Actual vs Predicted ESG Score:
ESG performance is expected to **improve** by 2.65 points.
NLP Insights
Sentiment of Description: Neutral (Score: -0.00)
Top Controversy Terms: ['Omniverse software', 'add-in board manufacturers', 'The Graphics segment']
AI-Generated ESG Insight: The company shows low ESG risk with minor concerns like networking segment comprises data center computing platforms, graphics segment offers geforce gpus.
Generate ESG PDF Report

9. Power BI Dashboard

9.1 Features & Layout

The Power BI dashboard for the ESG Risk Analyzer project was meticulously designed to provide a **comprehensive, at-a-glance view of ESG risks across 900+ companies**. The layout features:

- **KPI Cards**: Quick summary cards display total companies, average ESG risk score, total risk levels, and cumulative controversy scores.
- **Interactive Filters**: Users can filter companies based on Industry, Sector, and Controversy Level for a focused analysis.
- **ESG Risk Exposure Gauge**: A semi-circular meter visually highlights the overall ESG risk exposure in percentage terms.
- **Overall Company Table**: A detailed data table lists each company's Environmental, Social, and Governance scores, along with their ESG Risk Level and Controversy Level.
- **Top Companies Chart**: A horizontal bar graph showcases companies with the highest ESG scores.
- **Risk Level Progression**: A stacked bar chart tracks the distribution of companies across different ESG risk categories (Low, Medium, High, Negligible, Severe).

This streamlined layout ensures both a **macro** and **micro** view of ESG risks is readily accessible.

9.2 Sample Visualizations

Several critical visualizations are included in the dashboard to empower user-driven exploration:

- **Bar Charts**: Display top-performing and most at-risk companies based on ESG scores.
- **Risk Level Progression Chart**: Summarizes how many companies fall under each ESG risk category.
- **Dynamic KPI Indicators**: Update in real-time based on user-selected filters.
- **Interactive Table**: Allows users to scroll through detailed company-specific ESG score breakdowns and easily identify anomalies or patterns.

These visualizations ensure users can **interactively slice, drill down, and focus** on areas of concern without getting overwhelmed by raw numbers.

9.3 Insights Derived

From the dashboard, several actionable insights were observed:

- **Risk Distribution**: The majority of companies are concentrated in the Low and Medium risk categories, while Severe risk companies are very few.
- **Controversy Patterns**: Companies labelled with a High ESG Risk Level often also show a high or medium controversy level, suggesting a strong correlation.
- **Top Performers**: Companies like Adobe Inc., Essex Property Trust, and Aptiv Plc emerge as leaders with the lowest ESG scores.
- **Sector-Wise Differences**: Different industries show varying ESG risk concentrations, highlighting the need for sector-specific ESG strategies.

The Power BI dashboard transforms raw ESG data into **clear, strategic intelligence**, making it a powerful decision-support tool for analysts, investors, and sustainability teams.

10. Challenges Faced & Solutions

Building a comprehensive ESG Risk Analyzer involved overcoming multiple challenges across different phases of the project. Each hurdle was an opportunity to refine the approach and strengthen the solution:

1.Data Inconsistencies and Missing Values:
The initial datasets collected from multiple sources contained missing entries, inconsistent formats, and non-standardized categories (e.g., sector names, controversy levels).
Solution: Rigorous data cleaning was implemented using Pandas, ensuring all essential fields were imputed, standardized, or dropped thoughtfully without losing critical information.

2. Feature Alignment for Machine Learning Models:
During model training and prediction, mismatches between the feature columns in training data and real-time input data caused model failures.
Solution: Strict column alignment was enforced across the Random Forest and XGBoost pipelines, ensuring a consistent order and format for all features used during both training and inference.

3. NLP Challenges in Insight Generation:
While generating AI-based ESG insights using the distilled GPT-2 model, early results were often incomplete, repetitive, or vague.
Solution: Model prompts were carefully designed, and generation parameters (like temperature and max length) were fine-tuned. Additional post-processing was also applied to produce meaningful, structured 2-line insights for each company.

4. Controversy Keyword Extraction Errors:
When using RAKE for extracting top controversy terms, some company descriptions had missing or null values, leading to processing errors.
Solution: Robust exception handling was introduced, skipping empty descriptions and defaulting to placeholder keywords when necessary.

5. Dashboard Integration and Performance Issues:
Integrating large datasets into Streamlit and Power BI dashboards occasionally caused slow rendering and interaction delays.
Solution: Data aggregation techniques (e.g., pre-computed summaries) and efficient filtering mechanisms were implemented to optimize performance without sacrificing interactivity.

6. Model Evaluation and Comparison Complexity:
Comparing Random Forest and XGBoost involved tuning multiple hyperparameters and evaluating them across different metrics like accuracy, precision, recall, F1-score, MAE, and MSE, which was resource-intensive.
Solution: An organized evaluation pipeline was built, automating model comparison and allowing easy selection of the best model based on multiple criteria.

Challenge	Description	Solution
Data Inconsistencies	Missing values, inconsistent sector names, controversy levels, and non-standard formats in the dataset.	Rigorous data cleaning using Pandas with imputation, standardization, and careful dropping of records.
Feature Alignment	Mismatched features during ML model training and prediction caused model errors.	Enforced strict feature order and alignment across training and real-time prediction pipelines.
NLP Insight Generation Issues	DistilGPT-2 initially produced vague or incomplete ESG insights.	Tuned model prompts, adjusted generation parameters, and applied post-processing for better output.
Keyword Extraction Errors	Missing descriptions caused failures in RAKE and KeyBERT keyword extraction.	Implemented exception handling to skip or replace empty descriptions gracefully.
Dashboard Performance	Large datasets caused slow loading and interaction lags in Streamlit and Power BI dashboards.	Used pre-aggregation, optimized filtering, and lightweight UI components to enhance performance.
Model Comparison Complexity	Hyperparameter tuning and metric evaluation for Random Forest vs. XGBoost was time-consuming.	Built an automated model evaluation pipeline for efficient comparison and selection.

11. Conclusion

The ESG Risk Analyzer project represents a significant step toward making Environmental, Social, and Governance (ESG) risk assessment smarter, faster, and more accessible. In today's world, where sustainability and corporate responsibility are crucial, this tool empowers investors, analysts, and policymakers with data-driven, real-time ESG intelligence. Through the integration of machine learning models like Random Forest and XGBoost, natural language processing techniques including sentiment analysis and keyword extraction, and AI-generated summaries powered by distilGPT-2, the project creates a powerful end-to-end system for ESG analysis.

The combination of Streamlit-based dashboards for individual company analysis and Power BI dashboards for overall company comparisons ensures flexibility and depth in visual exploration. Users can filter companies based on sectors, industries, and controversy levels, view ESG risk score progressions, and access AI-driven insights for faster decision-making. This multi-platform visualization strategy ensures that users at different technical levels can effectively interact with the ESG risk data.

Throughout the journey, the project encountered several challenges, such as handling missing or inconsistent data, aligning features for machine learning models, and generating meaningful, concise AI insights. However, each challenge was systematically addressed using careful data preprocessing, robust feature engineering, model fine-tuning, and lightweight yet effective AI models. These solutions not only enhanced the reliability of the system but also significantly improved its overall interpretability and performance.

Ultimately, this project highlights how artificial intelligence, when thoughtfully applied, can elevate ESG risk analysis from static reports to dynamic, predictive, and insightful decision-making tools. The ESG Risk Analyzer stands as a scalable foundation upon which future enhancements—like real-time news integration, advanced forecasting, or sector-specific ESG benchmarking—can be built. It reaffirms the vital role of data-driven approaches in promoting sustainable investing and responsible corporate governance for a better future.

12. Future Enhancements

While the ESG Risk Analyzer provides a robust foundation for automated ESG risk assessment, there are numerous opportunities to expand its capabilities and impact. Future enhancements can further enrich its accuracy, usability, and practical relevance in the dynamic field of ESG analysis.

1. **Real-Time Data Integration**:
   Currently, the system operates on a static dataset. Future versions can integrate real-time ESG news feeds, regulatory updates, and sustainability disclosures through APIs. This would enable continuous risk assessment and more timely alerts for stakeholders.

2. **Advanced Sentiment Analysis**:
   While basic sentiment analysis using TextBlob has been implemented, the project can evolve by incorporating transformer-based sentiment models like BERT or RoBERTa for deeper contextual understanding of ESG-related narratives.

3. **Deep Learning Models for Risk Prediction**:
   Beyond Random Forest and XGBoost, future models could include deep neural networks (DNNs) or LSTM-based architectures to capture more complex patterns in ESG data and historical trends.

4. **Sector-Specific ESG Modelling**:
   Introducing customized ESG risk models tailored to specific industries (e.g., energy, finance, technology) could improve prediction precision and make recommendations more sector-relevant.

5. **Enhanced AI Insight Generation**:
   Currently, insights are generated using a distilled GPT-2 model. Future versions could fine-tune larger, domain-specific models (like finetuned GPT-3 or LLaMA) on ESG-specific corpora to produce even more nuanced, detailed summaries.

6. **Mobile App or Cloud Deployment**:
   Making the ESG Risk Analyzer available as a mobile app or on cloud platforms like AWS, Azure, or GCP would ensure broader accessibility and scalability.

7. **Predictive Forecasting of ESG Scores**:
   Implementing time series forecasting models could allow users to predict future ESG scores and anticipate potential risk trends based on historical data and emerging controversies.

8. **Explainable AI (XAI) Integration**:
   Adding explainability frameworks like SHAP or LIME would help users understand the key drivers behind each ESG risk prediction, increasing trust and transparency in the system.

13. References

1. **S&P Global** – ESG Scores and Data Sources
   Website: https://www.spglobal.com/esg/
2. **Nifty 50 Companies** – Corporate ESG Reports and Filings
   Website: https://www.nseindia.com/
3. **Hugging Face Transformers** – Pretrained Models like distilgpt2
   Website: https://huggingface.co/transformers/
4. **Scikit-Learn Documentation** – Machine Learning Models (Random Forest, XGBoost)
   Website: https://scikit-learn.org/stable/
5. **XGBoost Documentation** – Gradient Boosting Framework
   Website: https://xgboost.readthedocs.io/
6. **Streamlit Documentation** – Frontend Framework for Interactive Dashboards
   Website: https://docs.streamlit.io/
7. **TextBlob Documentation** – Sentiment Analysis Library
   Website: https://textblob.readthedocs.io/en/dev/
8. **RAKE (Rapid Automatic Keyword Extraction) Paper**
   Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Automatic Keyword Extraction from Individual Documents. In *Text Mining: Applications and Theory*.
9. **spaCy Documentation** – NLP Toolkit
   Website: https://spacy.io/
10. **Power BI Documentation** – Business Intelligence Dashboarding
    Website: https://learn.microsoft.com/en-us/power-bi/
11. **Sustainability and ESG Research Papers** – For Background and Context
    Various academic papers and whitepapers on sustainable finance, ESG investing, and corporate social responsibility (CSR).
12. Amel-Zadeh, A., & Serafeim, G. (2018). ESG investing: Practices, Progress and Challenges. *Journal of Applied Corporate Finance*, 30(2), 96-104. https://doi.org/10.1111/jacf.12234
13. Dass, N., Poursoltani, R., & Gupta, V. (2020). Using Natural Language Processing to Predict ESG Controversies. *Journal of Financial Data Science*, 2(4), 10–26. https://doi.org/10.3905/jfds.2020.1.015

----------x----------