How can AI and data make safe water quality a universal reality?

The better the question. The better the answer. The better the world works.

EY

Shape the future with confidence

2026 EY AI & Data Challenge: Guidance and Suggestions for Participants

Background

Welcome to the 2026 EY AI & Data Challenge! Held annually, the EY AI & Data Challenge gives thousands of early-career professionals and university students the opportunity to use data, Artificial Intelligence (AI) and computing technology to create solutions that address critical climate issues, building a more sustainable future for society and the planet. By joining this challenge, you will become part of a vital community engaged in activism through the use of AI to solve an important global sustainability issue.

The 2026 challenge focuses on predicting water quality and promoting access to clean and safe water supplies. According to the World Health Organization (WHO), in 2022, 73% of the global population had access to safely managed drinking water which means nearly 2.2 billion people did not have access to clean and safe water. The United Nations has warned that climate change is expected to increase water stress and exacerbate pollution through rising temperatures and extreme weather. They are further concerned that inadequate water quality monitoring could put the health and livelihoods of 4.8 billion people at risk by 2030. Thus, our ability to monitor water supplies and understand the environmental and climate drivers of water quality will enable better decision-making and help ensure clean water for everyone. This is the primary purpose of this data challenge… to understand the drivers of water quality and build AI models that could proactively support local water management.

Water quality is a fundamental concern with direct implications for human health, especially when it comes to potable water. Poor water quality can lead to waterborne diseases, long-term exposure to toxic substances, and increased treatment costs. Vulnerable populations—such as children, the elderly, and those relying on untreated sources—are particularly at risk. As urbanization, agriculture, and industrial activities intensify, water sources are increasingly exposed to pollutants that can compromise drinking water safety. Monitoring key parameters such as alkalinity, salinity and phosphorus is essential to ensure water remains safe, palatable, and compliant with health standards. Understanding and managing these parameters is also vital for climate resilience, as changing weather patterns can alter water chemistry and pollution levels. Investing in water quality monitoring ensures healthier communities, protects natural resources, and supports sustainable development. Water quality is not just an environmental metric—it is a cornerstone of public health and urban sustainability.

Challenge Goals

Though the topic of water quality has been widely studied and documented [1,2] there is a need for increased awareness and open-source models that address the drivers of water quality. The primary goal of the data challenge will be to develop a robust machine learning model capable of predicting water quality across various river locations in South Africa. In addition to accurate predictions, the model should also identify and emphasize the key factors that significantly influence water quality.

Participants will be provided with a dataset from the United Nations Environment Programme (UNEP) containing three water quality parameters: **total alkalinity, electrical conductance (EC), and dissolved reactive phosphorus (DRP)**.
The dataset (Figure 1) was collected over 5 years from rivers across South Africa. Each sample includes the geographic coordinates (latitude and longitude) of the sampling site, the date of collection, and the corresponding water quality measurements.

GEMS Station Number	Latitude	Longitude	River Name	Sample Date	Electrical Conductance (uS/cm)	Total Alkalinity (mg/L)	Dissolved Reactive Phosphorus (ug/L)
ZAF00001	-28.760833	17.730278	Orange River	1/2/11	555	129	10
ZAF00001	-28.760833	17.730278	Orange River	1/16/11	297	95	64
ZAF00001	-28.760833	17.730278	Orange River	1/30/11	231	85	33
ZAF00001	-28.760833	17.730278	Orange River	2/13/11	217	78	35
ZAF00001	-28.760833	17.730278	Orange River	2/27/11	225	83	31
ZAF00001	-28.760833	17.730278	Orange River	3/13/11	246	91	41
ZAF00001	-28.760833	17.730278	Orange River	3/27/11	294	100	34
ZAF00001	-28.760833	17.730278	Orange River	4/10/11	198	76	33

Figure 1. *An example of the training data including sample station information (name, latitude, longitude, river), sample date, and the values of three water quality parameters.*

Using this dataset, participants are expected to build a machine learning model to predict water quality parameters for a separate validation dataset, which includes locations from different regions not present in the training data. The challenge also encourages participants to explore feature importance and provide insights into the factors most strongly associated with variations in water quality.

Dataset Summary

Target datasets:

Below is a summary of the target parameter datasets used for this challenge. Each dataset includes three key water quality parameters: **total alkalinity, electrical conductance (EC), and dissolved reactive phosphorus (DRP)**. In addition, each dataset covers a time period from 2011 through 2015. More details about these datasets and water quality parameters are below.

Training Data:

- 27957 total samples
- 9319 unique measurements (location and date)
- 162 unique measurement locations (sample stations)
- 86 unique rivers

Validation Data:

- 600 total samples
- 200 unique measurements (location and date)
- 24 unique measurement locations (sample stations)
- 19 unique rivers

Water Quality Parameters:

Measuring water quality parameters is essential to protect human health, support aquatic ecosystems, and ensure water is suitable for its intended use. Each of the parameters presented in this challenge are influenced by environmental (e.g., land use) and/or climate factors (e.g., rainfall, temperature). Below is a description and statistical summary of the training data used for the challenge.

a) **Total alkalinity** in water quality is a measure of the water's ability to neutralize acids, measured in parts per million (ppm) or milligrams per liter (mg/L) of calcium carbonate (CaCO3). It acts as a pH buffer, preventing wide fluctuations in acidity or alkalinity, which is crucial for aquatic ecosystems and water treatment processes. For drinking water, the recommended range is generally 30 to 400 ppm, while ideal levels vary depending on the application. According to several global sources, an acceptable range for "good" water quality is between 20 mg/L and 200 mg/L. High alkalinity in South Africa rivers is driven by both natural geologic factors (e.g., weathering of carbonate rocks) and human activities such as mining, urbanization and agricultural runoff. Low alkalinity is driven by catchment geology (e.g., sandstone rock) and vegetation type. Figure 2 shows the distribution of total alkalinity data in the training dataset.

b) **Electrical conductance (EC)** in water quality is a measure of how well water conducts an electrical current, which is an indicator of the concentration of dissolved ions like salts and minerals. Higher conductivity typically signals more dissolved solids, which can mean pollution, but it is also influenced by natural factors like geology and seasons. Measuring EC is a general way to assess water quality and can be used to monitor pollution, saltwater intrusion, and overall water health. EC is measured in microsiemens per centimeter (uS/cm). According to several global sources, an acceptable range for "good" water quality is below 800 uS/cm. High salinity in South Africa rivers is driven by a combination of natural factors and human activity. Key drivers include the region's climate, geology, agriculture, mining and wastewater discharge. Figure 3 shows the distribution of EC data in the training dataset.

c) **Dissolved reactive phosphorous (DRP)** in water quality is a measure of phosphate levels which are essential nutrients for plant and animal life, but their excess in water can cause water quality issues. High levels of DRP are often coincident with algal blooms (eutrophication) that lead to murky water, odors, and the potential for harmful toxins. In addition, high levels of DRP lead to oxygen depletion in the water which can harm fish and other aquatic life. According to several global sources, an acceptable range for "good" water quality is below 100 ug/L. High DRP levels in South Africa rivers are often caused by human activities such as agricultural runoff, sewage and industrial wastewater. Figure 4 shows the distribution of DRP data in the training dataset.

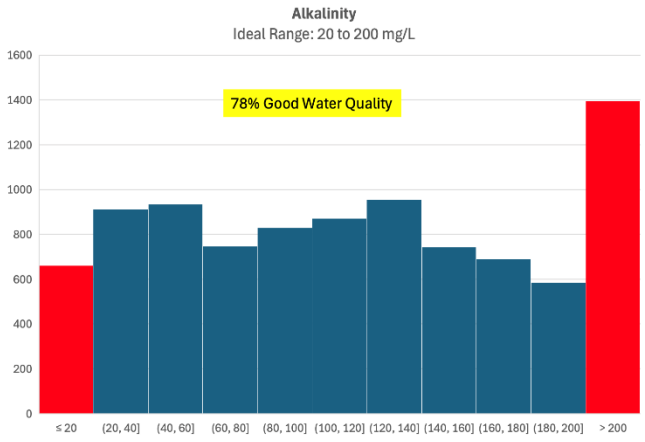

Figure 2. *A histogram of the total alkalinity parameter in the training data suggests a uniform distribution. Using a threshold of 20 to 200 mg/L, 78% of the water samples demonstrate "good" water quality.*

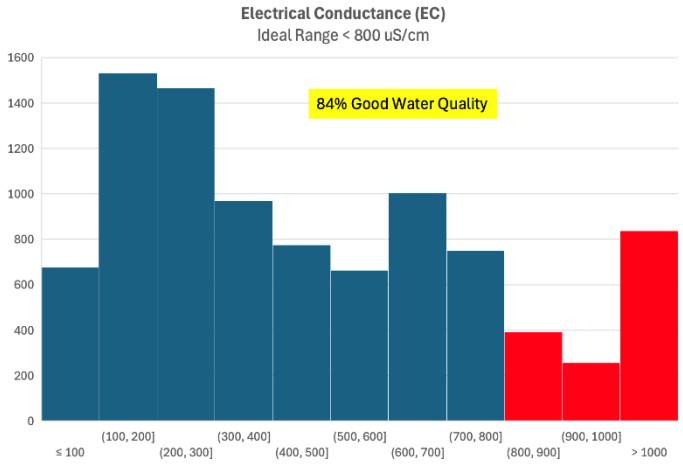

Figure 3. *A histogram of the EC parameter in the training data suggests a uniform distribution. Using a threshold of 800 uS/cm, 84% of the water samples demonstrate "good" water quality.*

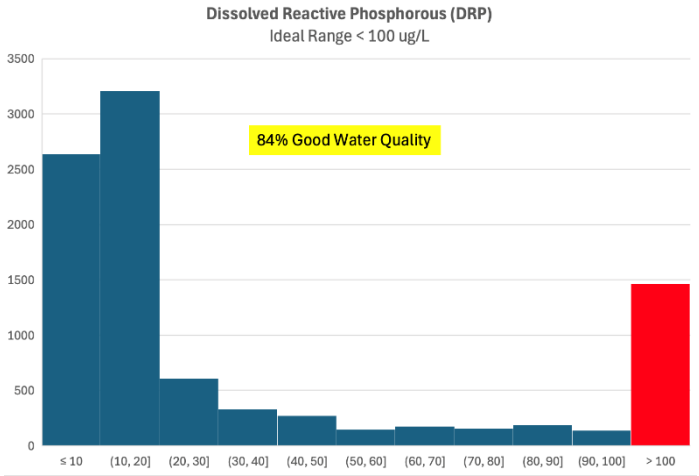

Figure 4. *A histogram of the DRP parameter in the training data suggests a non-symmetric bimodal distribution. Using a threshold of 100 ug/L, 84% of the water samples demonstrate "good" water quality.*

According to the United Nations Environment Programme (UNEP), based on data from 120 countries, only 56% of water bodies were rated as having "good" water quality. So, “good” water does not exist everywhere and is not available to everyone. An overall assessment of the training data used in this challenge yields similar results as 59% of the water samples are “good” water quality. To further illustrate the variability of water quality data used in this challenge each sample was tagged as poor water quality (value=0) or good water quality (value=1) using the thresholds presented in the prior section. If one of the parameters at a given sample location and date was classified as “poor” water quality, then the entire sample was considered “poor” water quality. This data was then used to calculate the mean water quality (0=poor, 1=good) at each sample location for all 5 years. Figure 5 shows the spatial distribution of water quality. Understanding the factors that are driving these spatial variations is one of the goals of this data challenge.

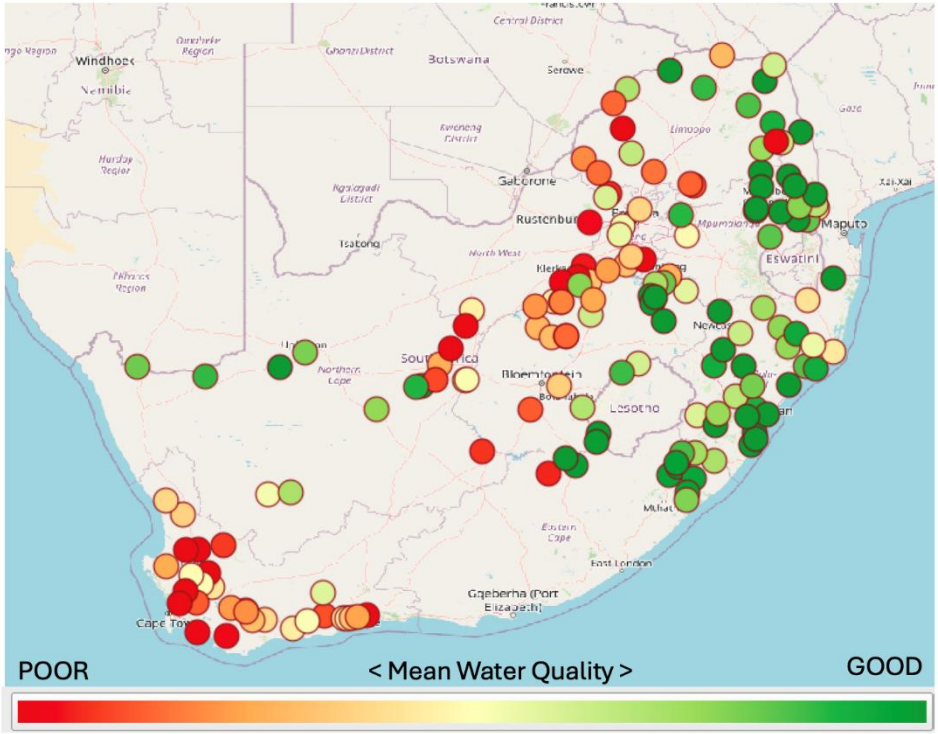

Figure 5. *The mean water quality (0=poor, 1=good) was calculated at each sample location over 5 years. Many regions exhibit a spatial bias in water quality. The areas near Cape Town, Pretoria and Johannesburg are the lowest mean water quality. The areas on the eastern shoreline near Durban and Kruger National Park are the highest mean water quality.*

Feature datasets:

Participants will be provided with a benchmark notebook that demonstrates the use of two publicly available satellite-based datasets (TerraClimate and Landsat). These datasets include features suitable to develop a baseline machine learning model to predict water quality variables. More details about these datasets are shown below. As a note, participants are allowed to use additional feature datasets for their models, provided those datasets are "open" and available to all public users and the source of such datasets are referenced in the models.

The launch of NASA's Landsat missions in 1999 and 2013 provides optical data at 30-meter spatial resolution and a revisit every 16 days with one mission and every eight days with two missions. This free and open data is readily available from the Microsoft Planetary Computer (https://planetarycomputer.microsoft.com/catalog). But optical data cannot penetrate clouds, so it is necessary to filter out clouds or select scenes that have very low levels of cloud cover. For this challenge, we have provided a sample Landsat Python notebook that reviews available scenes and demonstrates cloud filtering. This product can be used to assess the impacts of agriculture or urbanization on water quality. Below is an example Landsat cloud-filtered product showing the spatial variation of the Normalized Difference Vegetation Index (NDVI), which is a measure of vegetation density. The analysis region is centered in South Africa near two water sample stations along the Wilge River.

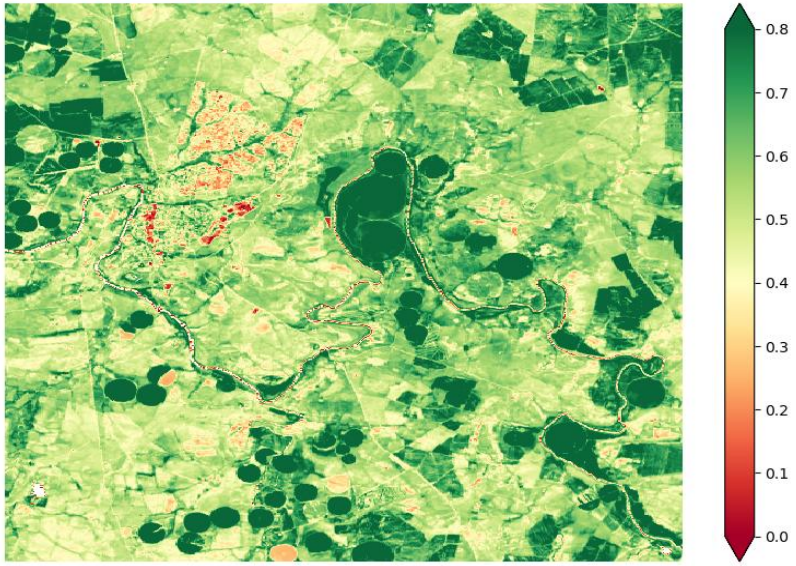

Figure 6. *Landsat Normalized Difference Vegetation Index (NDVI) over South Africa along the Wilge River. This product is based on a cloud-filtered scene from 04-Feb-2015. Areas of dark green are consistent with the presence of vegetation and agriculture croplands. Areas of dark red are consistent with dense urban environments or water. This information can be used in your digital model, as proximity to urbanization or agriculture can impact water quality. Credit: Brian Killough, EY (using data from NASA's Landsat mission on the Microsoft Planetary Computer).*

TerraClimate is a valuable satellite-based data source for global climate and water balance data. TerraClimate offers monthly data at a high spatial resolution of four kilometres, dating back to 1958. TerraClimate data includes **14 variables** which are essential for assessing environmental factors affecting water quality. There are six primary "measured" climate variables in the TerraClimate dataset. These include: maximum temperature (tmax), minimum temperature (tmin), vapor pressure (vap), precipitation accumulation (ppt), downward surface shortwave radiation (srad), and wind-speed (ws). The remaining eight variables are "derived" parameters and include: actual evapotranspiration (aet), reference evapotranspiration (pet), runoff (q), climate water deficit (def), soil moisture (soil), snow water equivalent (swe), Palmer Drought Severity Index (PDSI), and vapor pressure deficit (vpd).

One unique aspect of TerraClimate data is that it accounts for both climatic variations and hydrological balance. This provides valuable insights into water quality over time. Although TerraClimate data is available on various platforms, like Google Earth Engine, its integration with detailed topographic and climate-adjusted variables on the Microsoft Planetary Computer enhances usability for water quality modelling. In the context of water quality studies, TerraClimate data can serve as a foundational layer, helping researchers assess regions of poor water quality based on climate-driven parameters.

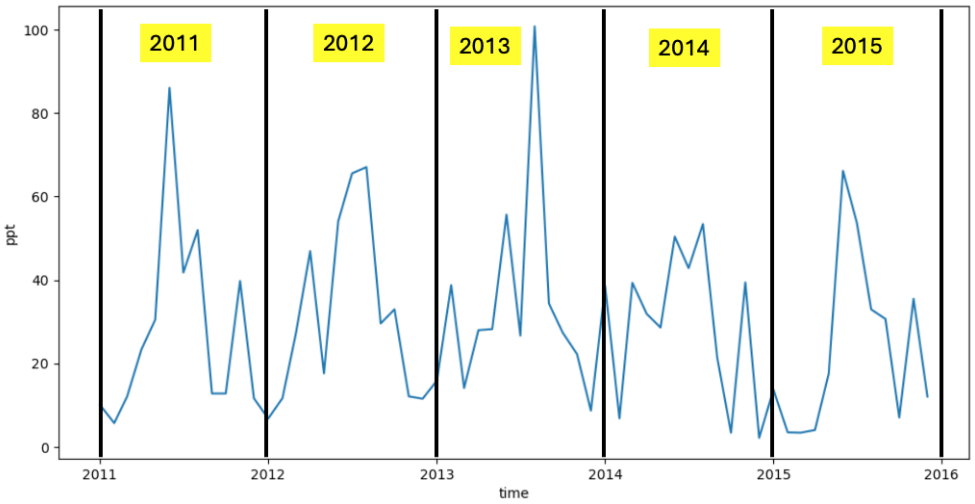

Figure 7. *Monthly accumulated precipitation from the TerraClimate dataset over a 400 square kilometer region near sample station #303 on the Bree River in South Africa. Seasonal variations in precipitation can be correlated with water quality parameters. Credit: Brian Killough, EY (using data from TerraClimate on the Microsoft Planetary Computer).*

Required Skills

Participants in this challenge can benefit from a basic understanding of statistics and Python programming, but there are no prerequisites for participation. Participating in this challenge will improve skills in machine learning, Artificial Intelligence (AI), data science, and working with satellite datasets. The data challenge has been designed to encourage beginners and those less familiar with AI and Python programming.

Computing Requirements

This data challenge was designed to run on a local computer with common computing resources (e.g., four cores, 32 GB memory). The configuration should include a Python programming environment and a code development tool (e.g., Jupyter). It is also possible to participate in this challenge using common cloud-based environments, such as those available from Microsoft (Azure), Google (Google Cloud, Earth Engine) or GitHub (Codespaces). As a new addition to the 2026 EY AI & Data Challenge, **Snowflake** is providing a powerful cloud computing environment to each participant. Participants can use this free environment for up to 120 days to develop their machine learning models. Use of this resource is **highly recommended** as it will significantly improve modelling performance and allow sharing and co-development among team members. In addition, participants will gain valuable experience using Snowflake's cloud environment which could enhance one's CV for future employment.

Please see the separate Snowflake resources links for more information about accessing this tool and associated training resources.

Model Development and Evaluation

Participants will develop machine learning / artificial intelligence (AI) models that can accurately predict water quality parameters (total alkalinity, electrical conductance, and dissolved reactive phosphorus) along rivers in South Africa, contributing to better environmental monitoring and public health outcomes. To get started, participants are provided with a sample benchmark Python notebook that will demonstrate a water quality prediction model.

This sample model is designed to use water quality data from the "target" dataset that was collected over five years from local rivers in South Africa. Landsat satellite data spectral bands and indices and TerraClimate variables are used as the "feature" datasets in the benchmark model. The model uses a common 70/30 training and testing split to evaluate model performance. The sample model produces an in-sample and out-of-sample R^2 score for each of the three water quality variables. This generalized model is then applied to a separate region in South Africa to produce a "submission" file with forecasted water quality parameters. This baseline submission produces a low mean R^2 score of 0.20 for the three water quality variables to allow significant improvements by data challenge participants.

Your task is to enhance these AI models and use them to predict the water quality parameters on the validation dataset, which include geolocations and dates from a different region in South Africa. The predictions must be saved in a single CSV file using the provided submission template and uploaded to the challenge platform. Your leaderboard score will be based on the mean R^2 score across all three parameters (total alkalinity, electrical conductance, and dissolved reactive phosphorus), which you are expected to significantly improve throughout the challenge.

Some suggestions for improved model performance include:

- Exploring additional Landsat spectral bands and spectral indices
- Exploring additional TerraClimate parameters
- Exploring spatial proximity to vegetation, water, and urbanization
- Exploring additional (publicly available) satellite and climate datasets
- Dataset and feature optimization for each water quality parameter
- Applying advanced data preprocessing techniques
- Experimenting with different regression algorithms (e.g., XGBoost)
- Perform hyperparameter tuning

As a note, participants are allowed to use additional datasets for their models, provided those datasets are "open" and available to all public users and the source of such datasets are referenced in the model.

Ten semi-finalist teams will be selected based on the highest R^2 score on the challenge leaderboard. The semi-finalists' models will be further reviewed for compliance, innovation and efficiency, with the highest scores resulting in the selection of up to five finalist teams.

Business Plan Development and Evaluation

Up to five finalist teams will be asked to develop a practical "business plan" that describes how their AI models could be applied by local beneficiaries to address local water quality decision-making. Finalists will be required to submit a written document (four pages or less) and a video (less than five minutes) that include the following: analysis approach, considerations for scaling such solutions to other locations, additional datasets that could improve model accuracy if given more time and resources, socioeconomic or health impact on vulnerable communities, and practical applications for local governments and policymakers. Participants should follow the provided template and use a strategic and well-structured approach while infusing creativity and considering generative-AI tools for completeness and enhanced impact.

Conclusions

The 2026 EY AI & Data Challenge is an excellent opportunity for students and young professionals to develop open-source solutions that can help bring clean, safe water to vulnerable communities. Entrants with top scores and compelling business plans will take home cash prizes and receive invitations to attend an exciting awards celebration. We look forward to seeing your results and wish you the best of luck.

References

1. Velibor Ilic; Maja Turk Sekulic; Maja Brboric; Jelena Radonic; Sonja Dmitrasinovic; Milan Stojkovic. Enhancing the monitoring system for river water quality: harnessing the power of satellite data and machine learning. Blue-Green Systems (2025) 7 (2): 338–352. https://doi.org/10.2166/bgs.2025.006

2. Maria Theresa Nakkazi; Albert Nkwasa; Analy Baltodano Martínez; Ann van Griensven. Linking land use and precipitation changes to water quality changes in Lake Victoria using earth observation data. Environmental Monitoring and Assessment (2024) 196:1104. https://doi.org/10.1007/s10661-024-13261-2

Questions?

Contact us at **datachallenge@ey.com**

EY | Building a better working world

EY is building a better working world by creating new value for clients, people, society and the planet, while building trust in capital markets.

Enabled by data, AI and advanced technology, EY teams help clients shape the future with confidence and develop answers for the most pressing issues of today and tomorrow.

EY teams work across a full spectrum of services in assurance, consulting, tax, strategy and transactions. Fueled by sector insights, a globally connected, multi-disciplinary network and diverse ecosystem partners, EY teams can provide services in more than 150 countries and territories.

All in to shape the future with confidence.

EYG no. 010440-25Gbl
ED None

ey.com