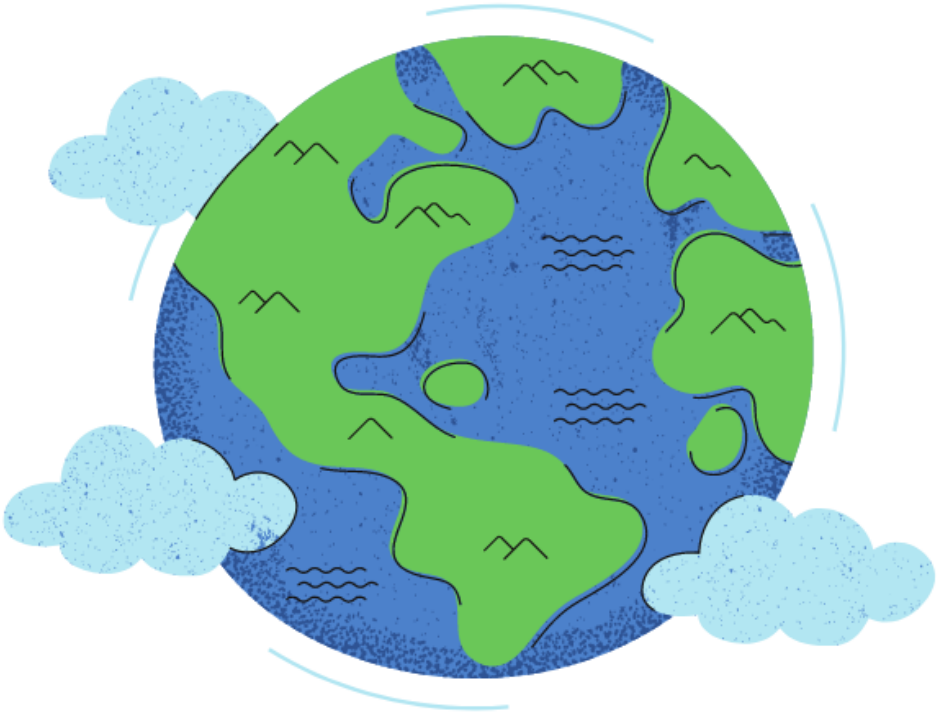
# SAUDI ARABIA WEATHER HISTORY



# HADOOP MAPREDUCE

1 DECEMBER, 2022

## BLACK BELTS


ABDULAZIZ OTAIF
MOHAMMED ALHADDAD
ELHAM ALGHAMDI
REEM ALSABTI
ESRAA ALZAHRANI
NORAH JADKARIM

## Table of Contents

# Introduction

## Problem Statement

Saudi Vision 2030 aims to achieve a renewable and sustainable energy supply of 9.5 GW by 2030. Therefore, the contribution of renewable energy to the overall energy mix will reach up to 50% of all energy supply in Saudi Arabia, which will contribute towards achieving sustainability and avoiding emissions and displacement of high-value fuel in electricity generation.
Renewable energy sources are natural forces that are strongly dependent on weather conditions. Therefore, bad weather conditions such as heavy clouds, rain, and sandstorms will reduce solar panels' energy supply significantly.
If we can predict these bad weather events before they happen, we can prepare hours, days, or even weeks in advance to balance energy supply sources and use traditional power sources to have enough energy supply to cover the demand.

## Goal

Our goal in this project is to leverage machine learning and data to predict weather conditions and various weather variables in the future. These predictions can be used to significantly increase the efficiency and reliability of renewable energy sources by estimating how much energy is likely to be produced from renewable energy technologies and how much is likely to be needed. These predictions can also be used and leveraged in many other areas such as agriculture where farmers can use weather forecasts to determine when to apply fertilizer and when to apply pesticides and insecticides and to avoid terrible damage to crops and soil erosion caused by unexpected bad weather. It can be used for safety purposes by providing citizens with actionable alerts and warnings about extreme weather events. There are many other areas where weather forecasting can be extremely beneficial such as aviation, retail, logistics and transportation, marine, and more.

# Dataset

## Overview

The dataset provides hourly weather recordings in the major cities and provinces of Saudi Arabia that were gathered from 2017 to 2019. The weather recordings in the dataset consist of cities where the recordings were obtained, the general weather conditions, information about the time and date of the recordings, and information about the measured weather metrics such as temperature, wind speed, humidity etc.

## Description

The dataset consists of 15 attributes and 249k+ instances. The attributes can be classified into three subcategories: the place of collection, the time and date of collection, and the collected weather condition and measured metrics.
Place of Collection:
**City**: consists of categorical values that represent the KSA standard names of the cities. It contains 13 *unique values* and no *missing values*.
Time & Date of Collection:
**Date:** represents the date on which the records were collected, as a date object. the date object is composed of year, month and day. It ranges from 01/01/2017 to 30/04/2019. It contains no *missing values*.
**Year:** A breakdown of the date attribute into years represented as numeric values in the format of *yyyy,* where *yyyy* represents the year (2017–2019). The attribute has no *missing values*.

**Month:** A breakdown of the date attribute into months represented as numeric values in the format of *m,* where *m* represents the month(1–12) sorted according to the Georgian calendar. The attribute has no *missing values*.
**Day:** A breakdown of the date attribute into days represented as numeric values in the format of *dd,* where *dd* represents the day (01–31). The attribute has no *missing values*.
**Time:** An attribute that represents the time of the collection represented as a time object consisting of the hours and minutes. this column has no *missing values*.
**Hour:** a breakdown of the time attribute into hours represented as numerical values in the format *hh*, where *hh* represents the hour of the day (01–24). this attribute has no *missing values.*
**Minute:** a breakdown of the time attribute into hours represented as numerical values in the format *mm*, where mm is supposed to represent the minutes(00–60), but instead it only contains 00 since the time attribute was approximated to the closest hour. this attribute has no *missing values.*
Weather Conditions & Metrics:
**Weather:** this attribute describes the general weather condition(clear, sunny, rainy .etc). It has 81 *unique values* and no *missing values*.
**Temp:** this attribute reflects the measurements of the temperature in Celsius. It has no *missing values.*
**Wind:** this attribute reflects the measurements of wind speed in km/h. It has no *missing values.*
**Humidity:** this attribute reflects the humidity percentage. It has 17 *missing values.*
**Barometer:** this attribute reflects the air pressure in mbar. It has no *missing values.*
**Visibility:** this attribute reflects the visibility distance in km**.** It has no *missing values.*

# Modeling Objectives

We have 4 main objectives:

1. The most frequent weather in each city
2. The most frequent wind speed in each city
3. The most frequent temperature in each city
4. The most frequent weather in each month.

How this project is related to Saudi's 2030 vision

- These objectives become doubly useful when applied to renewable energy sources. It is used to determine what is the most suitable renewable energy source for each city, and to predict how much power is likely to be produced and how much is likely to be needed.
- Renewable energy projects are one of the key drivers towards achieving **sustainability** that will contribute to avoiding emissions and the displacement of high-value fuel in electricity generation. **By 2030, the contribution of renewable energy to the overall energy mix will reach up to 50%.**
- Like wind turbines, the output of solar energy systems also depends largely on the weather—in this case, cloud cover. A recent study proposed a new method to use data from recently launched NASA satellites to predict the optical effects of clouds and the output of solar panels around the world.

# Preprocessing

The goal of this step was to transform the raw data into a compatible format for our task. To accomplish this goal we undertook several actions. First, we started by exploring the data. Second, we dropped unnecessary columns such as time, minute, and ID. Third, we converted humidity into floats by first removing the percent sign and then casting the values into floats. Subsequently, we dropped *NA values*. Finally, we mapped the 81 weather conditions into only 9 conditions. This was done by using substring matching to map the conditions. Then the remaining unmatched minority classes were dropped. For example, any weather condition that contained the substring cloud was mapped to Cloudy. Lastly, the final classes of weather conditions were Cloudy, Rain, Sunny, Duststorm، Fog, Clear, Haze, Sandstorm, and Overcast.

# Future Work

- Improve Hadoop MapReduce environment by:
  - Tuning performance of the CPU, memory, disk, and network.
  - Minimizing spilling by compression of mapper output.
  - Reuse jvm task.
  - Use Combine file input format for a bunch of smaller files.
- Use other aggregate functions such as min, max, and average to find more useful results.

# References


- https://www.vision2030.gov.sa/v2030/a-sustainable-saudi-vision/
- https://www.springwise.com/innovation/agriculture-energy/weather-forecasting-for-renewable-energy/
- https://theconversation.com/why-a-green-electricity-grid-depends-on-weather-forecasts-improving-152860
- https://innovation.engie.com/en/news/news/new-energies/AI-weather-forecast-optimize-renewable-energy/18235
- https://data-flair.training/blogs/mapreduce-performance-tuning/