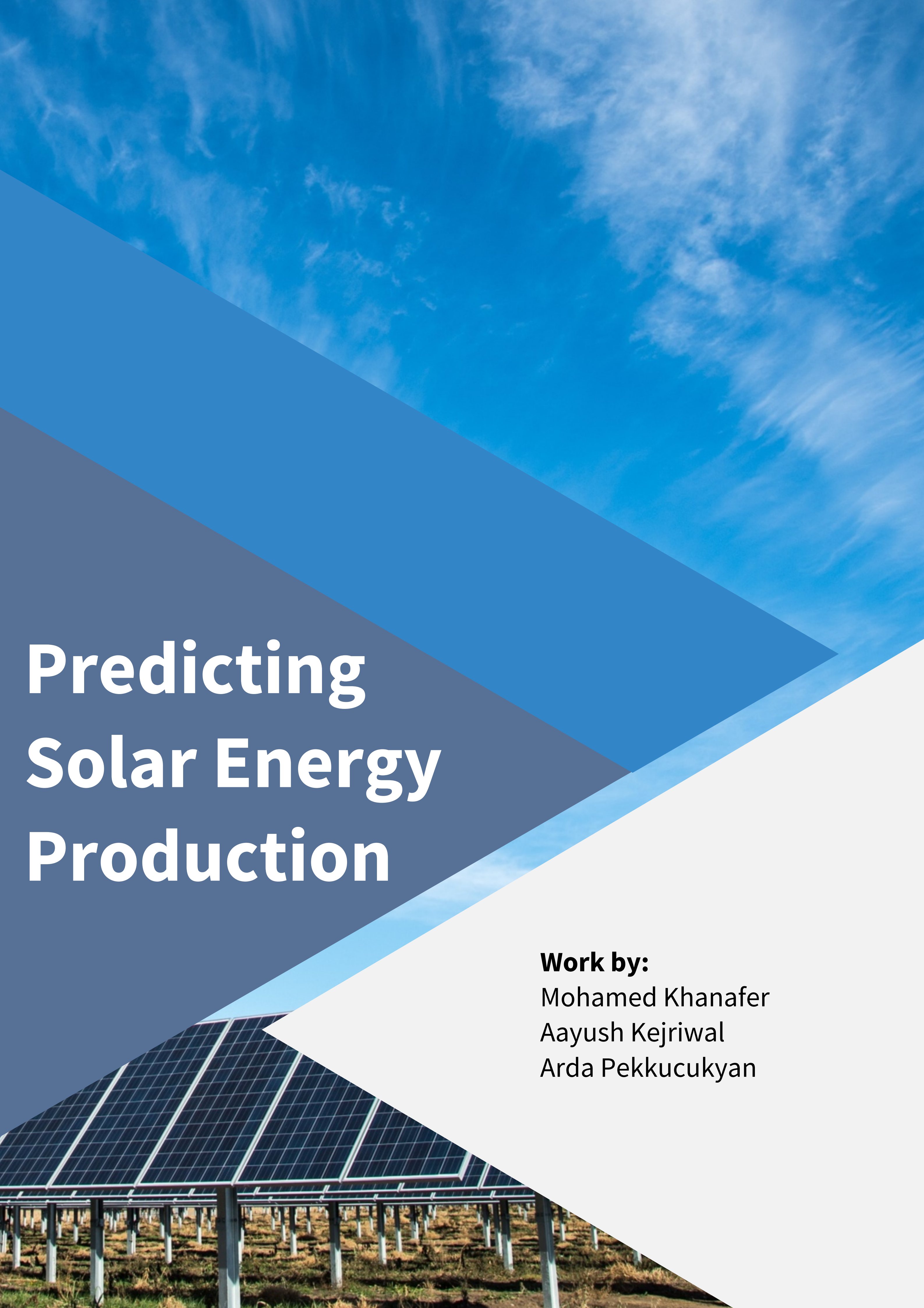

Predicting Solar Energy Production

Work by:
Mohamed Khanafer
Aayush Kejriwal
Arda Pekkucukyan

1. Objectives Summary

The goal of this project is to Train a Machine Learning model to try to predict solar energy production of 98 stations using the given dataset. This project is based on a Kaggle Competition:
https://www.kaggle.com/c/ams-2014-solar-energy-prediction-contest/overview.
We tried various approaches and the last model chosen is a SVM on the clustered stations' data with hyperparameter tuning.
Our model's score would be ranked in the top 20% of the competition. We were given a pre-processed dataset that we explain now.

2. The Dataset

In this project we work with a partly preprocessed dataset created from the original data given in the Kaggle dataset.
This dataset are found in the R object solar_dataset.RData, which contains a data.table with the following properties:

- A total dimension of 6909 rows and 456 columns;
- Each row corresponds to information of a particular day, ranging from 1994-01-01 to 2012-11-30. The first column, ‘Date’, informs you of which day corresponds to each row;
- The next 98 columns (from 2nd to 99th position) gives the real values of solar production recorded in 98 different weather stations. These columns are only informed until 2007-12-31 (row 5113); after this date these 98 columns contain NA or missing values. These missing values that must be predicted;
- The remaining columns are variables created from different weather predictors given in the Kaggle competition. They are the result of performing Principal Component Analysis, PCA, over the original data.

We are also provided with two other files:
1. station_info.csv: File with name, latitude, longitude, and elevation of each of the 98 stations.
2. additional_variables.rds: 100 new variables to optionally add to the ones in solar_contest.rds. All these variables correspond to real Numerical Weather Prediction, NWP, values. As in solar_contest.rds, each row corresponds to a particular day.
These are two auxiliary files, that we do not use in our analysis.