Inferences

1. Spacenet AI data : Calculate Number of building footprints and Forest Area Coverage by Image Segmentation. Prediction/Forecasting.
2. Countries -> MDS (clustering) -> select countries
3. Pick top 10 similar countries by using Similarity Search LSH
4. Create a profile of a country.
5. Use similarity search to find out the missing feature values of countries.

Hypothesis:

1. The dependency of developing countries on the developed countries reduces over the period.

target\_attributes = ['\_Urban\_population\_percent\_of\_total\_population',

'\_Net\_bilateral\_aid\_flows\_from\_DAC\_donors\_United\_States\_current\_US$',

'\_Net\_bilateral\_aid\_flows\_from\_DAC\_donors\_Total\_current\_US$',

'\_International\_tourism\_expenditures\_percent\_of\_total\_imports',

'\_Current\_Health\_Expenditure',

'\_Cost\_of\_business\_startup\_procedures\_percent\_of\_GNI\_per\_capita',

'\_Military\_expenditure\_percent\_of\_GDP',

'\_Research\_and\_development\_expenditure\_percent\_of\_GDP',

'\_Trade\_percent\_of\_GDP'

]

Set of features:

0. **Dependency:** Donors (US), Donors (Total), 7 other attributes from the following set of features.

1. **Economy Based:** GNI, new Business, domestic companies, cost of the business startup procedure
2. **Health Expenditure Based:** Current Health Expenditure, Prevalence in undernourishment, Hospital beds, Physicians, probability of dying among young ages
3. **Urbanization Core:** Urban Population Growth, Co2 Emission metric, Forest Area, Access to Electricity, Renewable Energy Consumption, Adjusted Saving: Co2 Damage, Adjusted Saving: Energy Depletion,
4. **Tourism:** International Tourism (departures), International Tourism (Arrivals), Tourism (expenditure), GDP, GNI
5. **Mortality:**  Urban Population Growth, infants death, Prevalence of Undernourishment, Pm2.5 air pollution, Prob. Of dying among youth, Current Health Expenditure, Physicians,
6. **RnD:** RnD Expenditure, Individuals using the internet, Compulsory education, Access to Electricity, People in RnD/1000,
7. **Trades:** Exports as a capacity of imports, Exports, Imports, Trade (% GDP), High Technology Export

The flow of the project:

1. Convert Unstructured to structured using TensorFlow. - Purva
   1. Code change & automate for the entire dataset
   2. Prediction values for unstructured data
   3. Prediction values for structured data
   4. Decide on the number of buildings
2. Find highly correlated features related to key attribute - Pulkit

* Min-max scaler
* User (country\_year to country\_year) mapping, to check the trend of one country1 in year x is linked to country2 in year y
* Predict the missing values
* Do the rescaling

1. Imputation of value by Similarity Search - Pulkit
2. Generate a Hypothesis tree to test all the hypotheses. - Pratik, Saurabh
   1. FInd bonferroni corrected values
3. Test the root Hypothesis. - Pratik, Saurabh
4. Recommend the changes for the coming years in the attributes to govern the required change in the root hypothesis. Show Rmse values - Purva, Saurabh
5. Visualization of our inferences - Pulkit
   1. Dashboard for selecting 5 variables from MDS and generating a realtime P-value, Basically I can remove PCP and add a button “find P value”, on that event the MDS updates so does P value of the dependent variable from dropdown. Also, I can see Similarity search of country and year wrt to other country and year and can show top 5 similar values.
   2. Create a short video for the presentation
6. Create Readme - Pulkit
7. Pratik Item-item , Pulkit user-user

Links

toDo : Add Tables and Figure. Limit each section to number of paragraphs given by Prof.

Project Report , add citations properly

<https://docs.google.com/document/d/1M4XDB7LjsN5m7SYOY9t62Z5Usv-IURa6L4mkQec4TFU/edit>

Pulkit part-1:

* We selected 57 highly correlated features Based on our SDG:11. While clicking on an attribute in MDS, changes Barchart, GeoMap and Stacked time series chart.
* Bar chart shows you the 10 most and least valued country for our selected attribute.
* Clicking on a country in WorldMap will stack it into time series chart so that you can compare the progress of countries.

Pulkit part-2:

* We can find similarity of a country’s growth in a year wrt to other countries. For Example: growth of Mexico in 2019 is similar to the growth of Mexico in 2018, TUR in 2019 and Poland in 2019.
* Let’s move onto more insights:
  + urban population growth in 2000 of US was 1.5, while for Russia it was -0.42, negative. Now if we see trends in 2020, US has dropped to 0.65, while russia has increased to 0.08. We tend to link urban population growth to decrease in forest area, but the trend shows that some developing countries like India and Russia, the % of forest area is increasing compared to 2000 while practicing sustainable urbanization.

Pratik part-2:

* As explained earlier we select an attribute and make a hypothesis. Here we are making a hypothesis on the dependent attribute cost of a business startup. We will select some features and generate a hypothesis that our target feature is dependent on these features, let's say we select high technology exports, new business density, mineral rents, RnD, and net bilateral flow. Now to verify if our hypothesis is correct or not we generate the P-value for the selected features and our dependent feature.
* As we can see that the p-value generated for these features is very very low and thus, we infer that our hypothesis is true. Also, we can see the weights of each of the attributes required for obtaining the dependent attribute. This is how we exhaustively generated hypothesis to create the hypothesis tree.

######################### Approx Presentation Break-up #########################

Purva: Start - Data Description

Pulkit: Architecture - Similarity Search

Pratik: Hypothesis Testing

Saurabh: Recommendation - Results

Pratik: Inference and Conclusion

######################## Section Timing Break-ups ############################

Goals to Data Description - 1 minute

Architechture - 1 minute

Methedology - 1 minute

Video - 2.5 minutes

Inferences and conclusion - 1 minute