**MM 226 – Materials Informatics**

**Assignment 2 Report**

**Name:** Shaksham K. Nigam

**Roll No.:** 230005045

• **Prompts Used for each AI**

The prompt which i used across four different AI models – **ChatGPT, Claude, Google Gemini, and Perplexity**:

“Generate a .csv file of 60 data points for Co,Co-Ni based alloys,while providing me with composition (chemical composition of alloy and the respective phases), processing condition (such as Heat treatments at specific case and Test

temperature),microstructural insights (grains size, phases, dislocation density),mechanical properties with units (YS, UTS, Hardness, Elongation, strain rate, strain hardening exponent and coefficient) while giving proper variability, missing values, and minor inconsistencies in the data fields"

**Chatgpt** produced a downloadable **csv** file, whereas **Gemini,Claude** generated a **csv file** which I had to **copy-paste** into excel and **Perplexity** gave me a **python code**, which upon running generated a csv file which I was able to download. All of this data was combined in one Google sheet named- Generated\_Data\_MM226

Link - Generated\_Data

• **Data Cleaning Methodology**

To clean the data, python codes were implemented for various inconsistencies and variabilities.

Python’s libraries-namely, Pandas, Numpy,etc were utilised for the purpose of cleaning the data set.

1. **Making the dislocation density column uniform for all the data sets.**  For the data sets produced by Gemini,Perplexity and Claude, the value contained scientific notation ‘E’, and it was made similar with the data produced by ChatGpT. ChatGpT had the column heading as 'Dislocation Density (×10¹⁴ m⁻²)',whereas the other data sets had 'Dislocation Density (m^2)', which were made consistent by utilising data[column] = data[column].apply(lambda x: x / 1e14 if pd.notnull(x) else x). The data files were then saved with a new name than the ones which were uploaded.

2. **Cleaning of the data**

A Function to clean non-numeric values in specified numeric columns was implemented to clean the numeric columns with any string values in it

“dataframe[column] = pd.to\_numeric(dataframe[column], errors='coerce')”

The pd.to\_numeric(value) replaced any non-numeric value with NaN, thus giving a clean numeric valued column.

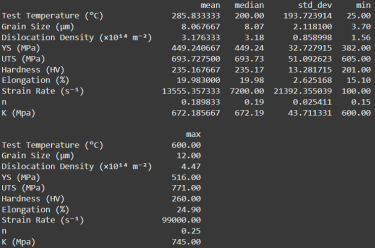
3. **Filling of Missing Values**

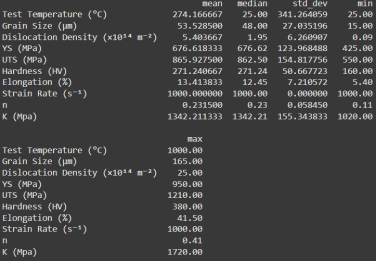
| **Method** | **Columns Affected** | **Justification** | **Reasoning** | **Benefits** |
| --- | --- | --- | --- | --- |
| **Forward Fill**  **(ffill())** | Phases,HeatTreatment, Test Temperature (°C) | Maintain  continuity  in non-numeric  data. | These  columns  typically have  consistent  values  unless  explicitly  changed. | Reduces errors by  assuming continuity; maintaining data  integrity. |
| **Mean Filling**  **(fillna(mean()))** | Grain Size, Dislocation Density, YS, UTS,  Hardness, Elongation, Strain Rate, n, K. | Use mean  to preserve data distribution and  avoid bias. | Mean is a  central  tendency  measure that  minimizes  impact on  data  distribution. | Preserves data  quality; maintains  statistical accuracy;  avoids introducing  bias. |

**4.Standardization of Data**

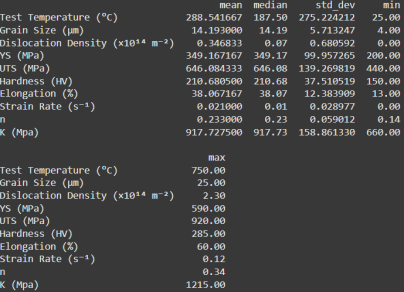
Round off function was used to standardize all of the data to 2 decimal places. Every numeric data column was standardized using this function

dataframe[column] = dataframe[column].apply(lambda x: round(x, 2) if pd.notnull(x) else x) • **Computed Mean,Median,Std\_Deviation,Min and Max of dataset 1.ChatGPT**

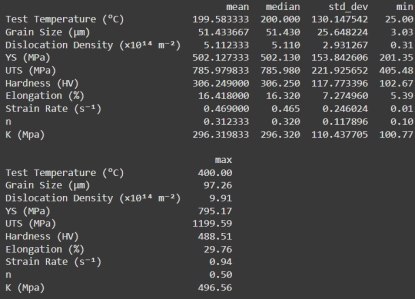
** 2.Claude**

****

**3.Gemini**

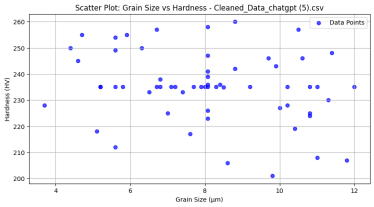
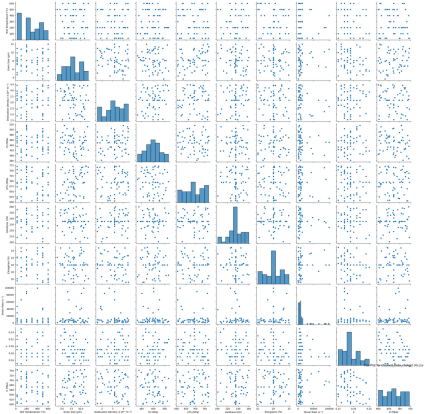
****

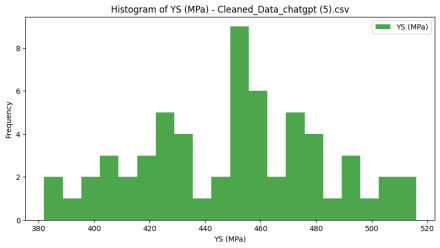
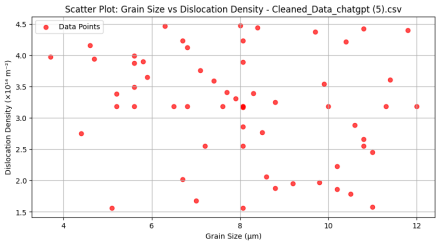
**4.Perplexity**

****

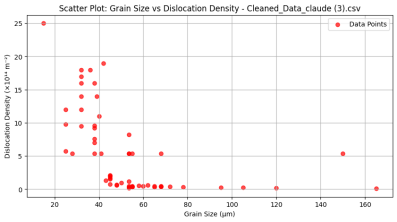
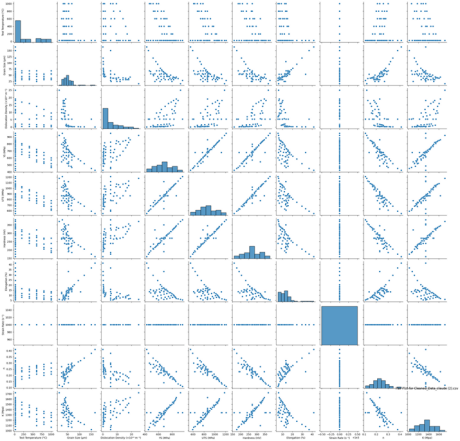
• **Comparative Plot Analysis of the AI Generated Data Sets**   **1.ChatGPT**

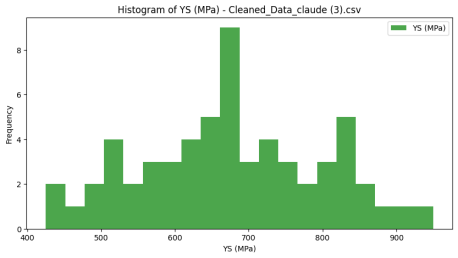
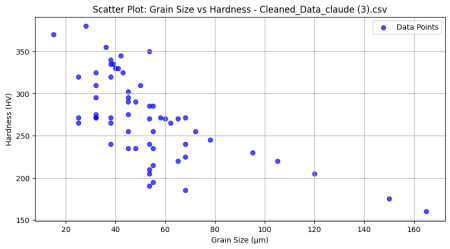
PairPlot of the properties



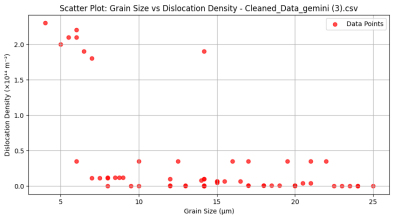
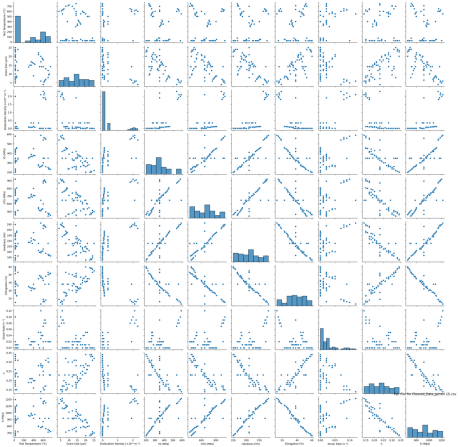
**2.Claude**

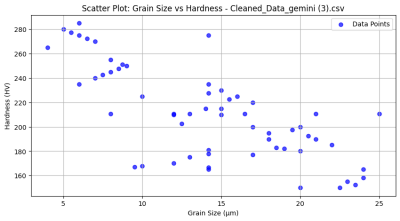
PairPlot of the properties



**3.Gemini**

PairPlot of the properties





**4.Perplexity**

PairPlot of the properties





**Comparative Analysis of the AI generated datasets**

| **Criteria** | **Claude** | **Gemini** | **Perplexity** | **ChatGPT** |
| --- | --- | --- | --- | --- |
| **Data** | High- | Plausible Co-Ni | Diverse | Simple Co-Ni |

| **Plausibility** | performance  alloys with  plausible  mechanical  properties. | based alloys  with typical  heat  treatments. | compositions  and phases;  some unusual  properties. | compositions  with plausible  mechanical  properties. |
| --- | --- | --- | --- | --- |
| **Consistency** | Consistent with high  performance  alloys. | Generally  consistent but  some high  elongation  values. | Wide variability; potential  inconsistencies due to broad  scope. | Consistent with simpler alloy  systems. |
| **Value Ranges** | Realistic ranges for UTS, YS,  and hardness. | Moderate  ranges; some  high elongation values. | Wide ranges;  includes  unusually high  values. | Narrower  ranges  compared to  others. |
| **Introduced**  **Inconsistencies** | Minimal  inconsistencies. | Some  inconsistencies in elongation  values. | More  inconsistencies due to diverse  data. | Few  inconsistencies. |
| **Discovered**  **Relationships** | Strong  correlations  between UTS,  YS, and  hardness. | Similar  correlations;  elongation less directly related. | Complex  relationships  due to diverse  data. | Correlations  consistent with simpler alloys. |

**Comparision with Manual DataSet**

**1. Elongation**

**2.Hardness**

****

**3.UTS**

****

**4.YS**

****

• **Comparative Analysis from the plots**

| Criteria | ChatGPT  Dataset | Claude  Dataset | Gemini  Dataset | Perplexity  Dataset | Manual Data |
| --- | --- | --- | --- | --- | --- |
| Ultimate  Tensile  Strength  (UTS) | Moderate  values (~700  MPa),  consistent with simpler alloys. | High values  (~850 MPa),  plausible for  high  performance alloys. | Lower values  (~650 MPa),  plausible but  limited in  scope. | Wide range  (~800  MPa),  includes  some  unusually  high  values. | Highest  values (~950 MPa),  consistent  with  experimental alloy  systems. |
| Yield  Strength  (YS) | Moderate  values (~450  MPa),  consistent with simpler alloys. | High values  (~650 MPa),  plausible for  high-strength alloys. | Lower values  (~350 MPa),  less consistent with expected  ranges. | Moderate  values  (~500  MPa), wide  variability  but  plausible. | High values  (~600 MPa),  consistent  with  experimental results. |
| Hardness  (HV) | Moderate  hardness  (~250 HV),  plausible for  simpler alloys. | High  hardness  (~275 HV),  consistent  with high  performance alloys. | Lower  hardness  (~225 HV),  plausible but  limited in  scope. | Wide range  (~300 HV),  includes  some  unusually  high  values. | Highest  hardness  (~375 HV),  consistent  with  experimental alloy  systems. |
| Elongation  (%) | Moderate  elongation  (~20%),  consistent with simpler alloys. | Low  elongation  (~15%),  typical for  high-strength alloys. | Higher  elongation  (~35%), less  common but  plausible for  certain alloys. | Wide range  (~25%),  includes  some  unusually  high  values. | Moderate  elongation  (~20%),  consistent  with  experimental alloy  systems. |

| Grain Size  (μm) | Smaller grain  size, typical for refined  microstructures in simpler  alloys. | Moderate  grain size,  consistent  with high  performance alloys. | Smaller grain  size, typical for refined  microstructures in certain  alloys. | Wide range  of grain  sizes,  includes  unusually  large  values. | Larger grain  size, typical  for coarse  grained  experimental systems. |
| --- | --- | --- | --- | --- | --- |

Link for all the data cleaning and analysis - Google Colab

**Brief Conclusion**

AI-generated synthetic datasets can effectively simulate complex material property relationships with sufficient realism and diversity. While these datasets **require careful cleaning** and human oversight, they are valuable tools for **training machine learning models** or simulating case studies in materials informatics, especially when real-world data is scarce or sensitive. The **Claude dataset** emerged as particularly promising for **predicting high-performance alloy** properties, while **ChatGPT** was suitable for **simpler systems**.**Perplexity dataset** allows **exploration of diverse systems** but requires careful oversight due to variability. **Manual data** remains the **gold standard for accuracy and realism**, serving as a benchmark for comparison.

This study highlights the potential of AI in augmenting materials science research by providing diverse and realistic synthetic data.

Drive link containing all resources - Assignment\_2