

53

Nous avons donc réalisé des **régressions** sur le dataset contenant les températures moyennes, les variations et les déviations des températures.

Heatmap des corrélations entre les variables

| Age Group | Percentage |
|-----------|------------|
| 18-24 | ~10% |
| 25-34 | ~35% |
| 35-44 | ~25% |
| 45-54 | ~20% |
| 55-64 | ~15% |
| 65-74 | ~10% |
| 75-84 | ~5% |
| 85+ | ~2% |

Heatmap showing the correlation matrix for the variables: Year, AvgTemperature, Standard Deviation, and Temperature change. The color scale ranges from -0.9 (dark purple) to 0.9 (dark red).

| | Year | AvgTemperature | Standard Deviation | Temperature change |
|--------------------|-------|----------------|--------------------|--------------------|
| Year | 1.0 | -0.15 | -0.1 | -0.1 |
| AvgTemperature | -0.15 | 1.0 | -0.2 | -0.1 |
| Standard Deviation | -0.1 | -0.2 | 1.0 | 0.85 |
| Temperature change | -0.1 | -0.1 | 0.85 | 1.0 |

Nous observons une corrélation de **-0.769915** entre les variables *Standard Deviation* et *AvgTemperature*.

Standard Deviation et *AvgTemperature* sont assez grandement **négativement corrélées**.

Lorsqu'une de ces deux variables augmente, l'autre diminue plus ou moins.

Les autres variables ne semblent pas vraiment corrélées.

Nous allons essayer de prédire la température moyenne en fonction des autres variables et en prenant en compte l'ordre chronologique de la colonne Year.

Nous avons ensuite testé **différents modèles, hyper-paramètres, variables et paramètres d'évaluation** pour essayer de trouver le modèle le plus pertinent.

Pour cela, nous avons **séparé les données initiales en deux parties**, une partie d'apprentissage qui contient 80% des données, et une autre partie qui représente les données de test et qui contient les 20% des données restantes.

Les modèles seront alors entraînés sur les données d'apprentissage puis évalués sur les données de test.

Nous pouvons voir ci-dessous que pour les deux modèles, la variable qui sert vraiment pour prédire la température moyenne est la *Standard Déviation*.

Ceci n'est pas surprenant, cette observation est cohérente avec les **analyses de corrélations** précédemment effectuées.

XGBoost qui avait l'air plus performant, prend beaucoup en compte pour sa prédiction, les variables *Temperature Change* et *Year* par rapport au modèle Random Forest.

Nous pouvons également constater que le pays qui sert le plus est la **Mongolie**, et cela peut s'expliquer par le fait que c'est le pays avec la température moyenne la plus basse au monde dans nos données d'après nos analyses de la partie **EDA - Température**.