|  |  |  |  |
| --- | --- | --- | --- |
| Partie | Texte dit à l’oral | Qui | Temps |
| Introduction | La recherche du bonheur est sans aucun doute l’une des plus grandes quêtes de l’humanité. Nous aspirons tous à atteindre cet état de bien-être durable.  La Data Analyse permet d’apporter des réponses à des questions essentielles :   * Quels sont les principaux facteurs sociaux, économiques et politiques qui influencent le bonheur à l’échelle mondiale ? * Comment ces facteurs ont-ils évolué au fil du temps en Europe ? * L’argent fait-il le bonheur ?   Dans cette présentation, vous découvrirez nos hypothèses, nos sources de données, des visualisations percutantes, des modélisations approfondies, ainsi que nos analyses et conclusions. | Patrick | 45s |
| Hypothèses | Pour répondre aux problématiques, nous avons émis des hypothèses générales au niveau mondial  Par exemple :  “La variable sociale, telles que l’accompagnement social et l’espérance de vie en bonne santé, jouent un rôle crucial dans la satisfaction de vie des individus”  Et des hypothèses spécifiques au niveau Européen  Par exemple :  “Les interactions entre les variables économiques (PIB, liberté économique), sociales (soutien social, santé) et politiques (perception de la corruption) expliquent en grande partie les variations de bonheur entre les pays européens.” | Patrick | 45 s |
| Sources | L’analyse exploratoire s’appuie sur une base de données initiales disponible sur le site worldhappiness (montrer le lien).  Cette base regroupe différentes enquêtes réalisées à travers le monde comme La note du bonheur (Life Ladder), le Pib par habitant ou encore le niveau de support social.  Pour enrichir la base de données initiale, nous avons intégré une nouvelle enquête disponible sur le site Cato Institute (montrer le lien)  Celle-ci note la liberté humaine sur ses aspects personnelles et économiques  Vous pouvez apercevoir ici (présenter df fusionnées) la DataFrame finale avec la fusion des 2 bases de données. | Patrick | 35s |
| Viz Monde :  Carte interactive | La dataviz est une pratique essentielle pour transformer des données brutes des enquêtes en représentations visuelles. Dans cette partie vous verrez des visualisations au niveau Mondial, puis d’autres au niveau Européen.  Voici une carte mondiale interactive. Cette première visualisation vous montre l’évolution de l’indice de bonheur par pays entre 2005 et 2023. | Patrick | 30 s |
| Viz Monde : HeatMap | Voici une heatmap des corrélations entre l’indice du bonheur et les différents indicateurs sociaux, économiques et politiques. Cette visualisation nous permet d’identifier rapidement les relations les plus fortes, qu’elles soient positives ou négatives, et de mieux comprendre les facteurs qui influencent le bonheur de manière significative. | Patrick | 30 s |
| Viz Monde : Boxplots | NB : Les données du social support ont été multipliées par 10 et celles de l’espérance de vie en bonne santé ont été divisées par 10 pour faciliter la comparaison entre les graphiques.  Graphique 3 : Boxplot   * Répartition des données assez équilibrée car la médiane est quasiment dans le milieu de chaque boite sauf celle de l’espérance de vie ; * Données très peu dispersées car les boites ne sont pas très grandes ; * Valeurs extrêmes dans le bas du graphique, principalement sur la variable du soutien social et sur celle de l’espérance de vie. | Lodavé | 60s |
| Viz Monde :  Moy Par Continent | Graphique 4 : Histogramme des variables par continent   * Les pays ayant le PIB par habitant moyen le plus élevé ont également la note de bonheur la plus élevée (Océanie et Europe). * Sauf dans le cas de l’Amérique du Nord et de l’Asie qui ont un PIB par habitant supérieur à celui de l’Amérique du Sud (3ème meilleur indice de bonheur moyen). | Lodavé | 50s |
| Viz Monde :  Evolution indice bonheur par continent | Graphique 5 : Evolution de l’indice de bonheur par continent   * La note du bonheur de l’Asie est aux alentours de 5. Causes probables : inégalités de richesse, instabilité géopolitique * La note de l’Amérique du Nord était basse en 2007 et depuis s’est moins stabilisée par rapport autres continents. Facteur explicatif : Impact très fort de la crise économique. | Lodavé | 50s |
| Viz Europe:  Evolution indice bonheur | Graphique 6 : Evolution de l‘indice du bonheur de 2006 à 2021 par sous-régions européennes   * Division en deux groupes après 2010 :   + D’un côté l’Europe du Nord et de l’Ouest : jouissent d’une situation économique plus stable, de systèmes protection sociale et de santé plus développés   + L’Europe du Sud et de l’Est : en cause beaucoup d’inégalités économiques, de forts niveaux de chômage et un manque de confiance envers les politiques.     Graphique 7 : Evolution des variables en Europe de 2006 à 2021     * Malgré les écarts entre les sous régions, l’Europe demeure un continent relativement stable. * La variable ayant le plus fluctué (hors indice du bonheur) est le soutien social. | Lodavé | 100s |
| ML  Intro | Pour la partie machine learning, notre variable cible est le life ladder la technique applicable est la régression.  Nous avons séparé nos données en un jeu d’entrainement et un jeu de test. | Gaelle | 10s |
| ML  Pré-traitement | Avant de commencer la partie modélisation les données doivent être traitées :   * Nous avons remplacé les données manquantes par la médiane par région et par année. * Les algorithmes de machine learning prennent en entrée uniquement des valeurs numériques nous avons donc encodé les données catégorielles afin de les transformer en variable numérique. * Nous avons également mis à la l’échelle nos données afin de pouvoir entrainer différents modèles.     Ces traitements s’entrainent sur le jeu d’entrainement puis s’appliquent sur le jeu d’entrainement et de test. | Gaelle | 30s |
| ML  Modèles | Nos données sont alors prêtes à être utilisé par les différents modèles de machine learning : nous avons choisi d’appliquer les modèles :   * De régression linéaire * Decision tree * Random forest * XGBoost     Pour chaque modèle qui le permet vous pouvez sélectionner les paramètres de profondeur et de nombre minimum d’échantillons par feuille. | Gaelle | 18s |
| ML  Decision tree  Je lance avec comme paramétrage 3 - 25 | Une des solutions permettant d'analyser la performance du modèle est de calculer des métriques:     * La MAE correspond à la moyenne des différences entre les vraies valeurs et les valeurs prédites. Donc plus proche de 0 est la valeur, meilleure est la note du modèle      * Le R² reflète la qualité de l'ajustement d'un modèle de régression. Plus proche de un est la valeur, meilleure est la note du modèle.     Un autre moyen de vérifier la qualité du modèle est le graphique de visualisation des prédictions. Moins le nuage de point est dispersé et donc plus proche de la ligne rouge meilleure est la prédiction.    Nous avons également déterminé pour chaque modèle les 4 variables les plus importantes. Nous remarquons l’impact important du PIB sur le life ladder. Les autres variables ont des taux faibles très proches.    Nous avons intégré un arbre de décision qui explique comment l’algorithme fonctionne. Il peut être interprété comme suit:   * Chaque nœud représente une décision basée sur une variable * Les valeurs dans les nœuds indiquent :   + squared erreur : l'erreur moyenne   + samples: nombre d'échantillons dans le nœud   + value: la valeur moyenne de l'indice de bonheur dans cette decision * Plus la couleur est foncée, plus le nombre d’échantillon est élevé | Gaelle | 2min |
| ML  Random forest  Appliquer sans les paramètres | Nous allons changer de modèle et appliquer le Random forest avec les paramètres par défaut.  Comme pour le decision tree les métriques sont aux alentours de 0.3 pour la MAE (train et test) et 0.8 pour le r² (train et test)  Le nuage de point parait un plus dispersé sur les prédictions d’un life ladder faible  L’importance des variables ressort pratiquement à l’identique que sur le modèle de Decision tree avec une variable essentielle : le PIB. Les 2 dernières variables ne sont pas les même mais leur taux est très proche  Pour le premier arbre de décision, nous pouvons remarquer que le modèle ne prend pas de décision de la même manière, en effet la deuxième condition que pose ici l’arbre concerne le positive affect et encore le PIB alors que sur le decision tree la deuxième condition concernait uniquement le PIB. | Gaelle | 1MIN |
| Analyse ML | Au moment de l'entrainement des modèles on remarque l'existence d'un surapprentissage. Pour limiter ce problème et optimiser le modèle afin qu'il soit le plus performant, nous avons utilisé un algorithme permettant d'optimiser les paramètres de chaque modèle.  Il convient alors de ré-entrainer les modèles avec ces paramètres optimisés et afficher les métriques MAE et R² pour chaque modèle.  R² : Random Forest est plus robuste avec un meilleur équilibre entre les ensembles train et test. XGBoost est très performant mais présente un léger risque de surajustement. Regression linéaire est un modèle robuste et simple. Decision Tree est fortement surajusté, ce qui en fait un mauvais choix sans régularisation ou assemblage.  Le ramdom forest a été créé pour limiter le sur apprentissage du decision tree.  Analyse des performances :     * **Régression linéaire** présente des métriques très proches entre entraînement et test. Ce qui est idéal pour éviter le surapprentissage. Ce modèle est donc à privilégier. * **Random Forest** offre des performances légèrement supérieures en R2 et en MAE tout en restant équilibré. * **XGBoost** sera privilégié si la priorité est la performance brute et que le surapprentissage peut être atténué par ajustement d'autres hyperparamètres. * **Decision Tree** affiche une grande différence entre les mesures d'entraînement et de test ce qui montre un fort surapprentissage, le rendant le moins fiable.     On peut rapidement remarquer que **le PIB** est la variable la plus importante du set d'après la Heatmap présentée et les features importances des modèles entrainés.  Les autres variables ont des résultats très proches entre elles, ce qui explique cet ordre d'importance différent. | Gaelle |  |
| Conclusion | Cette analyse met en évidence les principaux facteurs influençant le bonheur à travers le monde et en Europe. Les modèles prédictifs ont permis de mieux comprendre l'impact des variables sociales, économiques et politiques sur le bien-être global. ... etc. |  | 1MIN20 |