Data Analytics

The Impact of Remote Work on Employee Mental Health

KOUROUMA Banawata

Octobre, 2024

Table of content

1. **Introduction ..... p 2**
   - Project Title
   - Objective/Goal
   - Plan

2. **Data Collection & Data Sources..... p 4**
   - API
   - Flat Files
   - Database Integration

3. **Data Cleaning & Exploratory Data Analysis..... p 6**

4. **SQL Integration..... p 10**
   - Database Structure (ERD)
   - SQL Queries & Insights

5. **API Exposure..... p 17**

6. **Machine Learning..... p 21**
   - Models Used
   - Feature Engineering

7. **Conclusion..... p 24**

8. **GDPR Compliance..... p 25**

9. **References..... p 26**

Introduction

Business Use Case:

In today's evolving work landscape, companies are increasingly adopting flexible work policies. However, many organizations still lack clarity on how these environments impact employee well-being and productivity.

This analysis aims to fill that gap by helping businesses to:

- **Identify Key Factors**: Determine which work environments are most conducive to reducing stress and improving productivity.
- **Inform Policy Changes**: Provide data-driven insights to refine remote and hybrid work policies in a way that supports employee mental health.
- **Enhance Employee Well-being**: Use the findings to offer personalized mental health support tailored to employees' roles and work locations.

Goal:

The goal of this project is to provide a comprehensive analysis of how different work environments—remote, onsite, and hybrid—impact employee mental health and productivity. The analysis focuses on the following key metrics:

- **Stress Levels**: Assess how stress levels vary across different work locations and job roles.
- **Productivity Change**: Measure productivity shifts during remote work and compare them to onsite and hybrid settings.
- **Work-Life Balance**: Investigate how different work environments affect employees' work-life balance, a critical factor in overall well-being.
- **Mental Health Conditions**: Explore the prevalence of mental health issues like anxiety, burnout, and depression across various work environments.

PLAN:

The project involves the following steps:

1. Data Collection: The dataset sourced from Kaggle includes demographics, work environments, and mental health indicators.

2. Data Cleaning and Preprocessing: Handling null values, standardizing data, and preparing for analysis.

3. Exploratory Data Analysis (EDA): Identifying relationships between work location and mental health/productivity indicators.

4. SQL Integration: Storing cleaned data in a MySQL database, running queries to extract key insights.

5. Visualization: Creating dashboards and visual representations in Tableau for better decision-making.

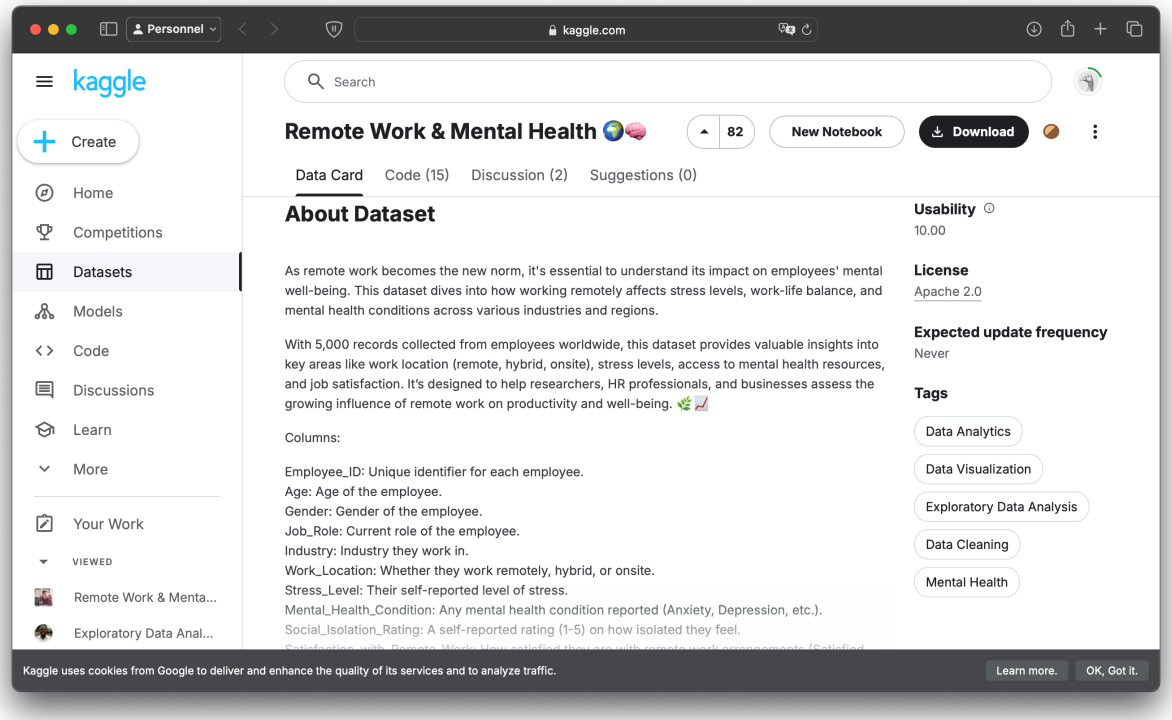
6. API Integration: Exposing data via Flask API to allow access to filtered mental health and productivity data. This will include endpoints, pagination, and potential nested filters.

7. Machine Learning : Building predictive models to identify at-risk employees based on work location and other factors.

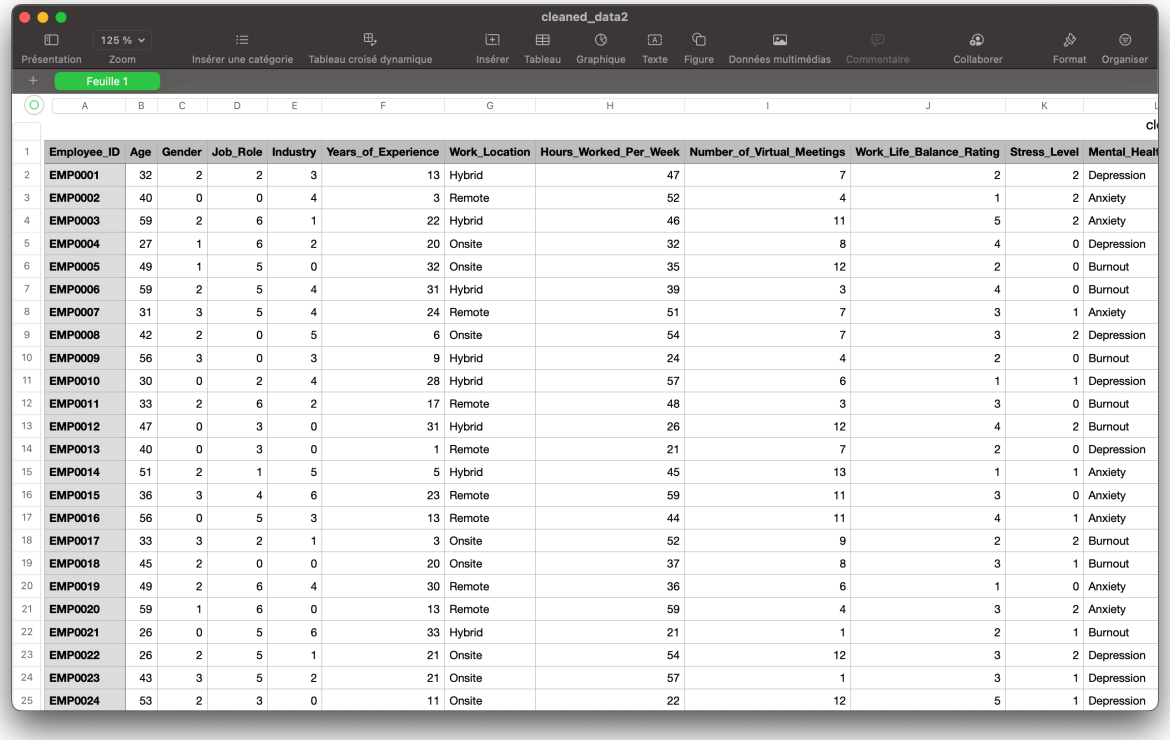
Data and data sources

This section covers the details of the Kaggle dataset used in the project, as well as API integration and database storage.

- **Kaggle Dataset**:
  The project utilized the **"Remote Work and Mental Health"** dataset from Kaggle. This dataset includes key employee information such as demographics, job roles, work environments (remote, onsite, hybrid), and mental health conditions (e.g., anxiety, burnout, depression). The dataset provided the foundational data for analysis and insights on how different work setups affect employee mental health and productivity.



- **API Integration**:
  To extend the data's usefulness, an API was set up. Although the primary data source was a cleaned flat file, the API allowed for interaction with additional external resources related to mental health and productivity. The API offered options for filtering data and included pagination to manage large datasets efficiently.



cleaned_data2

Employee_ID	Age	Gender	Job_Role	Industry	Years_of_Experience	Work_Location	Hours_Worked_Per_Week	Number_of_Virtual_Meetings	Work_Life_Balance_Rating	Stress_Level	Mental_Healt
EMP0001	32	2	2	3	13	Hybrid	47	7	2	2	Depression
EMP0002	40	0	0	4	3	Remote	52	4	1	2	Anxiety
EMP0003	59	2	6	1	22	Hybrid	46	11	5	2	Anxiety
EMP0004	27	1	6	2	20	Onsite	32	8	4	0	Depression
EMP0005	49	1	5	0	32	Onsite	35	12	2	0	Burnout
EMP0006	59	2	5	4	31	Hybrid	39	3	4	0	Burnout
EMP0007	31	3	5	4	24	Remote	51	7	3	1	Anxiety
EMP0008	42	2	0	5	6	Onsite	54	7	3	2	Depression
EMP0009	56	3	0	3	9	Hybrid	24	4	2	0	Burnout
EMP0010	30	0	2	4	28	Hybrid	57	6	1	1	Depression
EMP0011	33	2	6	2	17	Remote	48	3	3	0	Burnout
EMP0012	47	0	3	0	31	Hybrid	26	12	4	2	Burnout
EMP0013	40	0	3	0	1	Remote	21	7	2	0	Depression
EMP0014	51	2	1	5	5	Hybrid	45	13	1	1	Anxiety
EMP0015	36	3	4	6	23	Remote	59	11	3	0	Anxiety
EMP0016	56	0	5	3	13	Remote	44	11	4	1	Anxiety
EMP0017	33	3	2	1	3	Onsite	52	9	2	2	Burnout
EMP0018	45	2	0	0	20	Onsite	37	8	3	1	Burnout
EMP0019	49	2	6	4	30	Remote	36	6	1	0	Anxiety
EMP0020	59	1	6	0	13	Remote	59	4	3	2	Anxiety
EMP0021	26	0	5	6	33	Hybrid	21	1	2	1	Burnout
EMP0022	26	2	5	1	21	Onsite	54	12	3	2	Depression
EMP0023	43	3	5	2	21	Onsite	57	1	3	1	Depression
EMP0024	53	2	3	0	11	Onsite	22	12	5	1	Depression

- **Flat Files and MySQL Database**:
  The cleaned dataset was stored in a MySQL database using **SQLAlchemy** for efficient querying and management. This structured approach allowed for seamless data access and analysis through SQL queries. An **ERD (Entity-Relationship Diagram)** was also developed to visualize the database's structure and the relationships between entities, such as employees, job roles, departments, and work environments. This design helped guide the data analysis process.

Data collection

Kaggle Dataset Selection: The dataset titled "Remote Work and Mental Health" was sourced from Kaggle. This dataset was chosen for its relevance in analyzing the impact of remote, hybrid, and onsite work environments on employee well-being, including factors such as stress levels, work-life balance, and productivity. The dataset contains a wide range of variables such as job roles, mental health conditions, hours worked, and employee demographics, making it suitable for the scope of this analysis.
(Refer to the Kaggle dataset screenshot in the Data and Data Sources section.)

API Setup: While the main dataset was sourced from a flat file (CSV), an API was set up to enable real-time access to the mental health data. This API offered additional functionality by allowing for filtering, displaying, and analyzing the data in a more interactive manner. Although the flat file was the primary source of data, the API enhanced the analysis by providing easy access to the dataset and offering data-driven insights in real-time.
(Refer to the API integration screenshot in the API Integration section.)

Data Validation and Preparation: The collected data was reviewed and validated to ensure accuracy and relevance to the project's goals. This process involved downloading the CSV file, performing initial validation checks, and understanding the structure of the dataset to ensure that it included all necessary variables. After validation, the data was cleaned and stored in a MySQL database for further analysis and modeling.
(Refer to the flat file screenshot in the Data and Data Sources section.)

Data cleaning and Exploratory data analysis

Data Cleaning:

During this stage, null values were handled, and inconsistencies were resolved. For instance, null values in columns such as Department and Region were replaced with "Unknown" to maintain data integrity. Redundant entries were eliminated, and categorical variables were standardized for analysis.

Raw Data Before Cleaning: The initial dataset consisted of various columns, including employee demographics, work location, work-life balance, mental health conditions, and more. At this stage, the dataset contained both categorical and numeric variables. The raw data had not yet undergone any preprocessing or cleaning. This view shows the diversity of information available but also indicates that some preprocessing was required before further analysis.

	Employee_ID	Age	Gender	Job_Role	Industry	Years_of_Experience	Work_Location	Hours_Worked_Per_Week	Number_of_Virtual_Meetings	Wo
0	EMP0001	32	Non-binary	HR	Healthcare	13	Hybrid	47	7	
1	EMP0002	40	Female	Data Scientist	IT	3	Remote	52	4	
2	EMP0003	59	Non-binary	Software Engineer	Education	22	Hybrid	46	11	
3	EMP0004	27	Male	Software Engineer	Finance	20	Onsite	32	8	
4	EMP0005	49	Male	Sales	Consulting	32	Onsite	35	12	

Handling Missing Values: A critical aspect of the data cleaning process was identifying and managing missing values. Initially, the dataset exhibited a number of missing entries, particularly in columns like Mental_Health_Condition and Physical_Activity. This prompted the decision to handle missing values through appropriate imputation techniques, ensuring the dataset remained useful for analysis without compromising data quality.

```
Missing Values in each column:
Employee_ID                            0
Age                                    0
Gender                                 0
Job_Role                               0
Industry                               0
Years_of_Experience                    0
Work_Location                          0
Hours_Worked_Per_Week                  0
Number_of_Virtual_Meetings             0
Work_Life_Balance_Rating               0
Stress_Level                           0
Mental_Health_Condition             1196
Access_to_Mental_Health_Resources      0
Productivity_Change                    0
Social_Isolation_Rating                0
Satisfaction_with_Remote_Work          0
Company_Support_for_Remote_Work        0
Physical_Activity                   1629
Sleep_Quality                          0
Region                                 0
dtype: int64
Data Types:
Employee_ID                         object
Age                                  int64
Gender                              object
Job_Role                            object
Industry                            object
Years_of_Experience                  int64
Work_Location                       object
Hours_Worked_Per_Week                int64
Number_of_Virtual_Meetings           int64
Work_Life_Balance_Rating             int64
Stress_Level                        object
Mental_Health_Condition             object
Access_to_Mental_Health_Resources   object
Productivity_Change                 object
Social_Isolation_Rating              int64
Satisfaction_with_Remote_Work       object
Company_Support_for_Remote_Work      int64
Physical_Activity                   object
Sleep_Quality                       object
Region                              object
dtype: object
```

Missing Values After Imputation: After handling missing values, we re-evaluated the dataset to ensure all previously missing entries were accounted for. This view illustrates that the imputation process successfully filled in gaps across various columns, particularly in critical fields such as Mental_Health_Condition and Physical_Activity. The dataset was now clean, complete, and ready for further preprocessing and analysis.

```
Missing Values after Imputation:
Employee_ID                          0
Age                                  0
Gender                               0
Job_Role                             0
Industry                             0
Years_of_Experience                  0
Work_Location                        0
Hours_Worked_Per_Week                0
Number_of_Virtual_Meetings           0
Work_Life_Balance_Rating             0
Stress_Level                         0
Mental_Health_Condition              0
Access_to_Mental_Health_Resources    0
Productivity_Change                  0
Social_Isolation_Rating              0
Satisfaction_with_Remote_Work        0
Company_Support_for_Remote_Work      0
Physical_Activity                    0
Sleep_Quality                        0
Region                               0
dtype: int64
```

Dataset After Encoding: To prepare the dataset for machine learning models, we encoded categorical variables such as Job_Role, Work_Location, and Gender. This transformation converted these non-numeric fields into numeric formats, ensuring compatibility with machine learning algorithms. The screenshot reflects the dataset after encoding, where each categorical feature has been replaced with corresponding numeric values, facilitating smooth model training.

Dataset after Encoding:

	Employee_ID	Age	Gender	Job_Role	Industry	Years_of_Experience	Work_Location	Hours_Worked_Per_Week	Number_of_Virtual_Meetings	Work_Life_B
0	EMP0001	32	2	2	3	13	Hybrid	47	7	
1	EMP0002	40	0	0	4	3	Remote	52	4	
2	EMP0003	59	2	6	1	22	Hybrid	46	11	
3	EMP0004	27	1	6	2	20	Onsite	32	8	
4	EMP0005	49	1	5	0	32	Onsite	35	12	

Final Cleaned Dataset: The final cleaned dataset incorporates all preprocessing steps, including the handling of missing values and the encoding of categorical variables. At this stage, the dataset is ready for analysis, with key columns such as Work_Location, Hours_Worked_Per_Week, and Stress_Level prepared for deeper exploration in the EDA and machine learning stages.

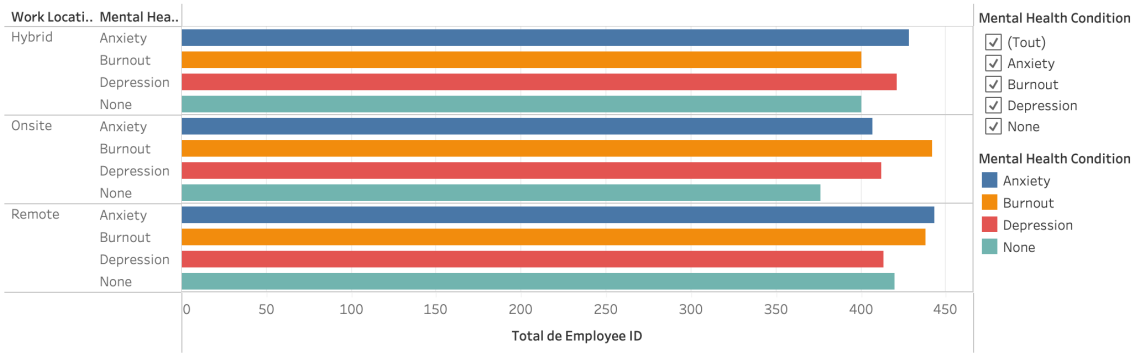
```
Work Location Unique Values: ['Hybrid' 'Remote' 'Onsite']
Mental Health Condition Unique Values: ['Depression' 'Anxiety' 'Burnout']
Mental Health Condition Counts:
Mental_Health_Condition
Burnout       2476
Anxiety       1278
Depression    1246
Name: count, dtype: int64
Stress Level Counts:
Stress_Level
0    1686
2    1669
1    1645
Name: count, dtype: int64
```

EDA:

Exploratory Data Analysis revealed significant relationships between **work location** and **mental health indicators** like stress levels and productivity change. Remote workers generally experienced less stress, while hybrid workers were more likely to experience higher stress and productivity changes.

Mental Health Conditions by Work Location:

Explanation: This bar chart shows the distribution of mental health conditions (Anxiety, Burnout, Depression, None) across different work locations (Hybrid, Remote, Onsite). It highlights that workers in hybrid environments experience higher instances of burnout and anxiety compared to other work locations. This suggests that the hybrid work environment may introduce additional stressors leading to poor mental health outcomes.

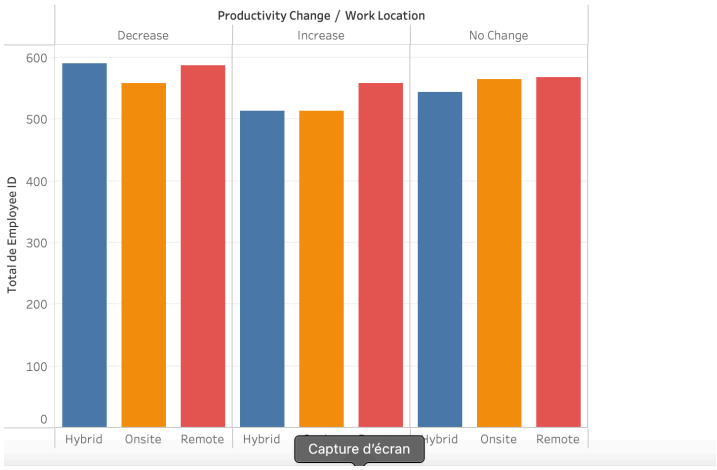


Stress Levels by Job Role:

Explanation: This bar chart displays stress levels (low, medium, high) across various job roles. The graph indicates that stress levels are fairly evenly distributed across roles, but certain roles, such as Software Engineer and Sales, tend to have higher instances of medium to high stress levels. Understanding this variation helps target stress management programs to specific roles.

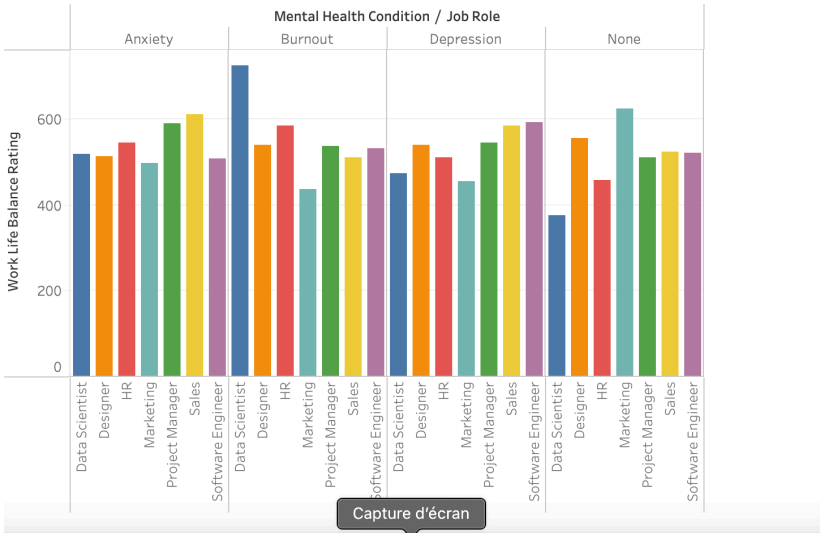
Productivity Change by Work Location:

Explanation: The stacked bar chart represents productivity changes (increase, decrease, no change) by work location. The chart shows that workers in onsite roles are more likely to experience productivity increases, while remote workers show a more balanced distribution between increased and decreased productivity. This suggests that the work environment plays a significant role in productivity outcomes.



Work-Life Balance by Job Role:

Explanation: This box plot compares work-life balance ratings across different job roles. It reveals that certain roles, like Data Scientist and Software Engineer, tend to have a better work-life balance compared to others, like Sales or Project Managers, where balance appears to be more strained. This could inform tailored recommendations for improving work-life balance based on job roles.



Insights from the EDA informed the next stages of the analysis, particularly the machine learning models and SQL queries.

Database type selection

MySQL Selection:

The decision to use MySQL as the database system was based on its efficient handling of structured data. MySQL, when combined with SQLAlchemy, offered the flexibility needed for querying and managing large datasets, which was critical for the project's objectives. It allowed for the seamless execution of SQL scripts, enabling deep analysis into key areas such as stress levels, work locations, and mental health conditions.

Integration with SQLAlchemy:

SQLAlchemy provided ORM (Object-Relational Mapping) functionality, ensuring that queries could be executed efficiently while maintaining scalability for future integrations. The flexibility of SQLAlchemy also opens the possibility for expanding the project to include additional data sources or API integration.

In this project, SQL was employed to store, manipulate, and analyze the data stored in a MySQL database. The cleaned dataset was transferred into MySQL, where several key queries were run to extract valuable insights regarding mental health, work location, and productivity. These queries provided a foundation for deeper analysis and visualization.

Key SQL Queries & Insights:

Stress Levels by Work Location:

```sql
SELECT rw.Work_Location, rw.Stress_Level, COUNT(*) AS num_employees
FROM remote_work_mental_health rw
GROUP BY rw.Work_Location, rw.Stress_Level;
```

This query groups employees by work location and stress level, showing how stress levels vary across different work environments.

Hybrid workers tend to report medium to high stress levels, while remote workers experience lower stress levels overall.

Mental Health Conditions by Work Location:

```sql
SELECT rw.Work_Location, rw.Mental_Health_Condition, jr.Job_Role_Name, COUNT(*)
FROM remote_work_mental_health rw
JOIN job_roles jr ON rw.Job_Role_ID = jr.Job_Role_ID
GROUP BY rw.Work_Location, rw.Mental_Health_Condition, jr.Job_Role_Name;
```

This query shows how mental health conditions (burnout, anxiety, etc.) are distributed across different work locations.

Hybrid workers have higher reports of anxiety and burnout compared to onsite and remote workers.

Productivity Change by Work Location:

```sql
SELECT rw.Work_Location, rw.Productivity_Change, COUNT(*) AS num_employees
FROM remote_work_mental_health rw
GROUP BY rw.Work_Location, rw.Productivity_Change;
```

This query displays the number of employees experiencing productivity changes based on their work location.

Remote workers tend to show a positive increase in productivity, while hybrid workers are more split between no change and a decrease in productivity.

Work-Life Balance by Job Role:

```sql
SELECT rw.Work_Location, rw.Work_Life_Balance_Rating, jr.Job_Role_Name, COUNT(*)
FROM remote_work_mental_health rw
JOIN job_roles jr ON rw.Job_Role_ID = jr.Job_Role_ID
GROUP BY rw.Work_Location, rw.Work_Life_Balance_Rating, jr.Job_Role_Name;
```

This query calculates the average work-life balance rating for each job role.
Software engineers report the highest work-life balance, while salespeople report lower work-life balance ratings.

Access to Mental Health Resources by Work Location:

```sql
SELECT Work_Location, Access_to_Mental_Health_Resources, COUNT(*) AS num_
FROM remote_work_mental_health
GROUP BY Work_Location, Access_to_Mental_Health_Resources;
```

This query identifies how access to mental health resources is distributed across work locations.
Remote workers have better access to mental health resources compared to onsite workers.

The integration of MySQL and SQLAlchemy proved crucial for this project, allowing efficient storage, querying, and analysis of the dataset. By leveraging SQL queries, we uncovered valuable insights regarding the relationship between work location and mental health outcomes. The use of SQL facilitated in-depth analysis of stress levels, work-life balance, and productivity, contributing significantly to the overall understanding of how different work environments impact employee well-being. These insights were further strengthened by visualizations and machine learning models, driving actionable business recommendations.

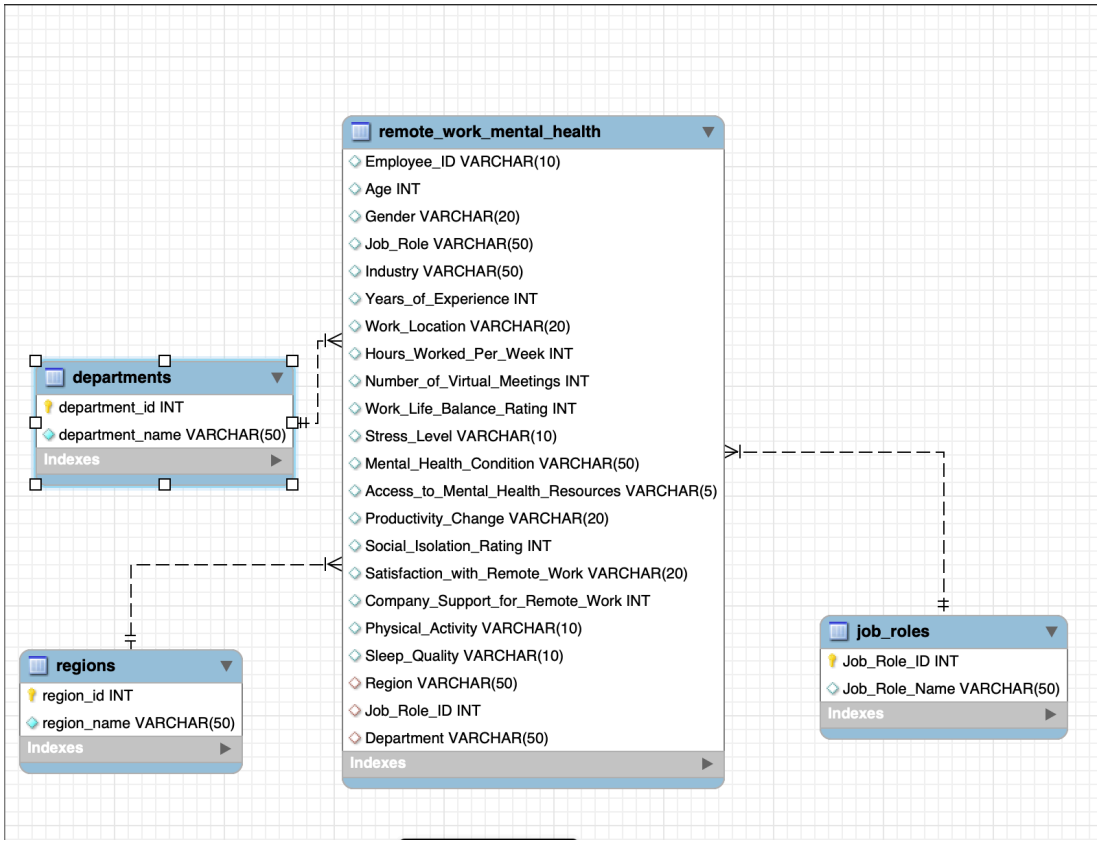
Entity Relationship Diagram (ERD):

The **ERD (Entity Relationship Diagram)** visually represents the structure of the MySQL database used in this project. It illustrates the relationships between the key entities, including **employees, job roles, departments, and regions**, providing a clear map of how these elements are interconnected.

Tables and Foreign Keys: The ERD outlines the relationships and constraints between tables using foreign keys, ensuring data consistency and integrity. Each table is logically linked—for example, employees are connected to job roles, departments, and regions, allowing us to analyze various factors that influence mental health and productivity.

Relevance to the Analysis: By structuring the database in this way, the ERD enables **structured queries** that can examine the relationships between different work environments and their impact on **mental health outcomes**, such as stress levels and productivity changes. It provides a foundation for **deep analytical insights**, allowing us to link specific job roles or regions with higher stress levels, better work-life balance, or changes in productivity.

The relationships and constraints illustrated in the ERD form the backbone of our database queries, ensuring accurate and meaningful results during analysis.



The established relationships between the tables allow for comprehensive querying, offering insights into how different work environments (remote, onsite, hybrid) and job roles affect mental health outcomes. By linking the *job_roles* and *remote_work_mental_health* tables, for instance, we can easily analyze how stress levels or productivity changes vary across different job categories, providing actionable insights for HR and management.

Similarly, the connection between *regions* and *departments* offers the opportunity to identify whether specific geographic areas or organizational units are more prone to mental health challenges, helping tailor wellness programs to regional or departmental needs.

API

API Integration:

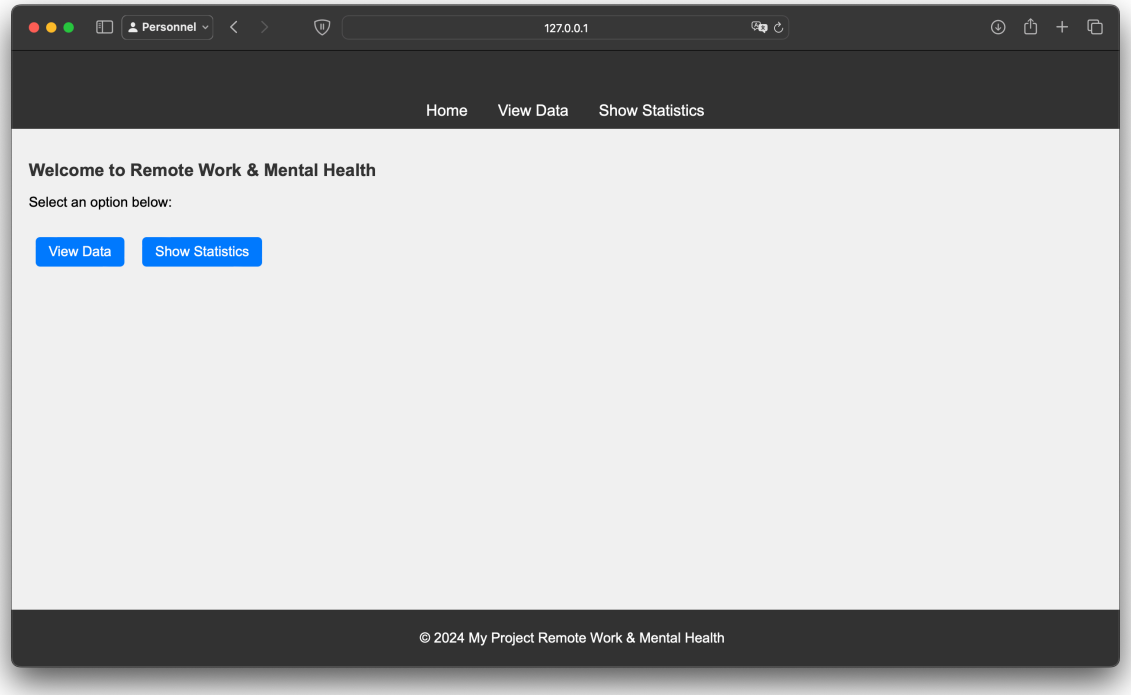
The API was built using Flask to serve the processed data and visual insights related to remote work and mental health. It reads data from a preprocessed CSV file (cleaned_data2.csv), which contains the cleaned and structured dataset derived from the original Kaggle dataset used throughout the analysis. This dataset includes variables such as employee job role, stress levels, work location, and productivity change.

The API exposes endpoints that allow users to view the dataset and access key statistical insights, providing a foundation for future scalability and potential interactive features.

Key Endpoints:

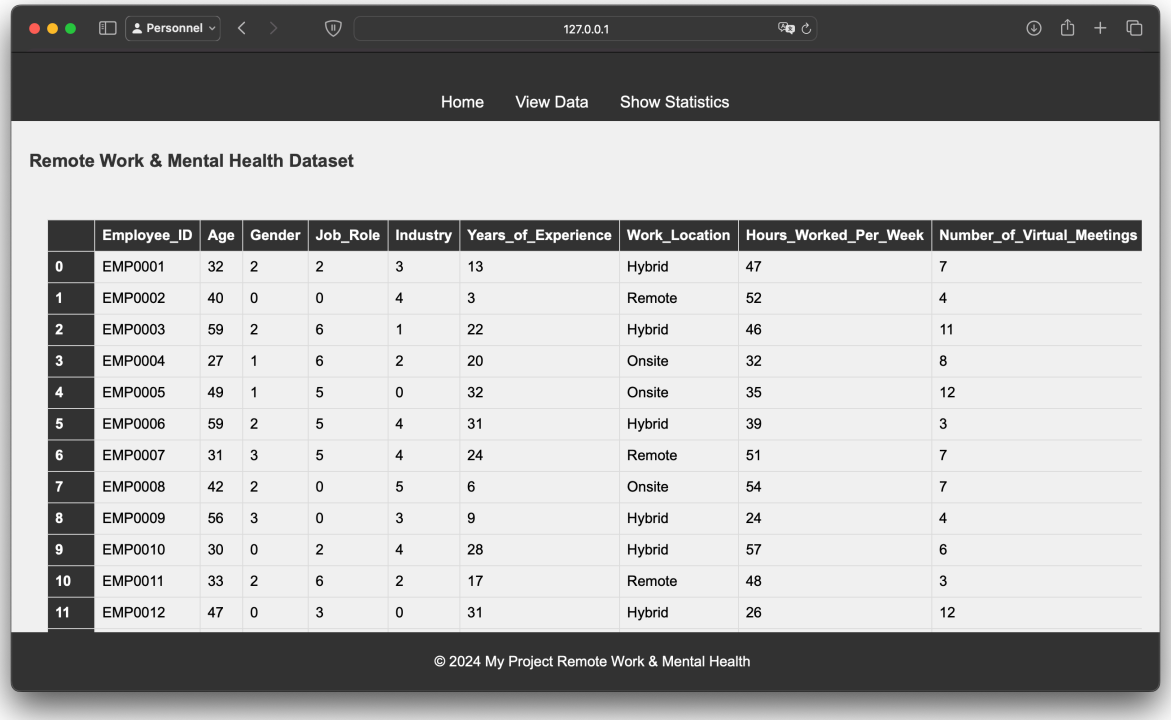
- **Home Route (/)**:

Purpose: The screen where users can choose between viewing data or seeing statistics.
Functionality: This is the entry point for the API where users are presented with two buttons: "View Data" and "Show Statistics." From here, users can navigate to the desired data interaction.



View Data Route (/view):

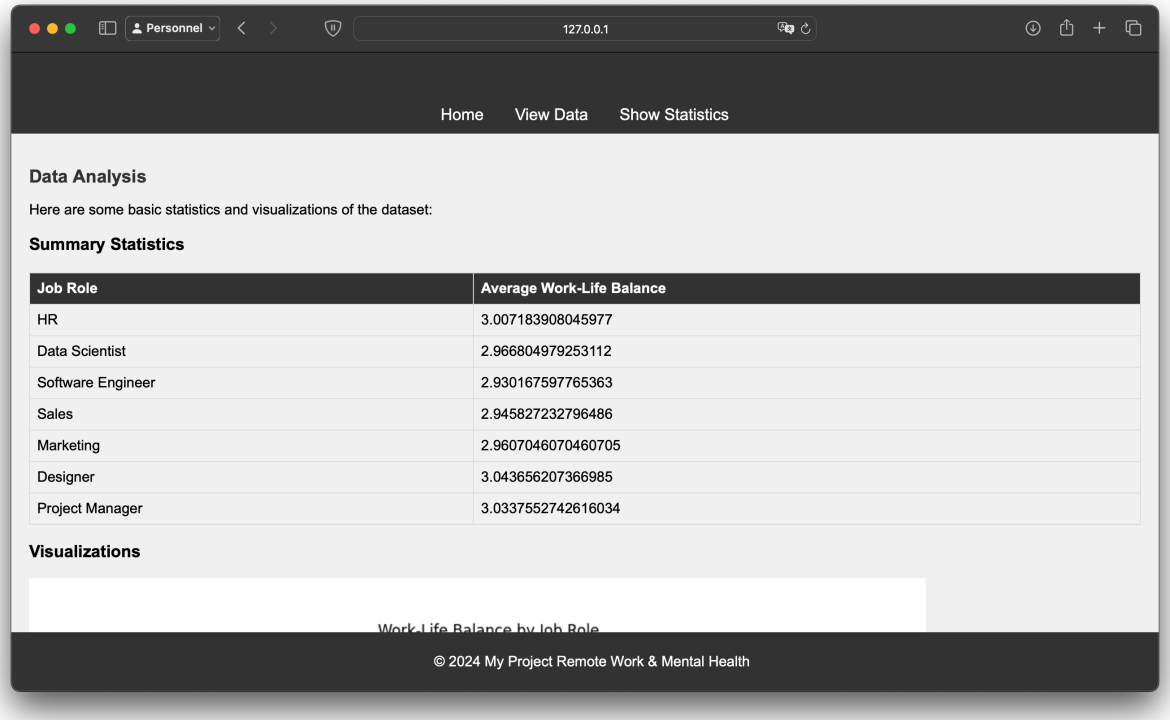
Purpose: This route enables users to visualize and interact with the dataset directly from the CSV file.
Functionality: The user can explore the dataset by scrolling through the records in a neatly formatted table. This enhances accessibility to the raw data, making it easy to review.

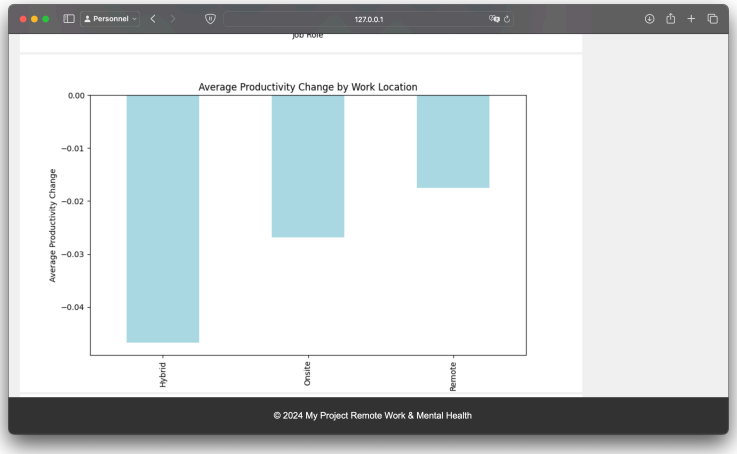


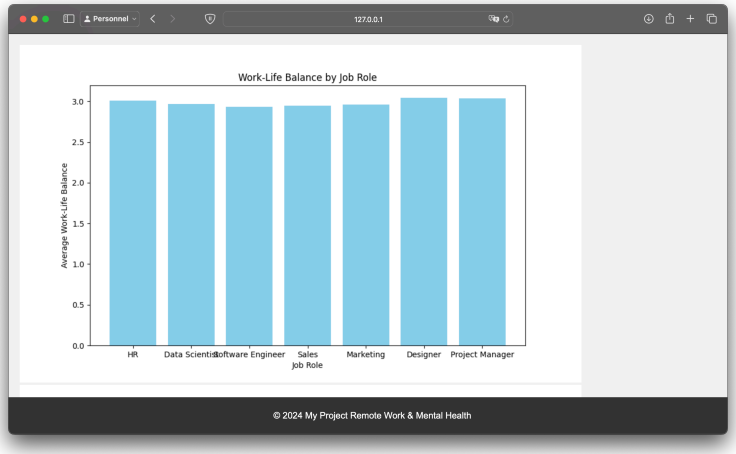
	Employee_ID	Age	Gender	Job_Role	Industry	Years_of_Experience	Work_Location	Hours_Worked_Per_Week	Number_of_Virtual_Meetings
0	EMP0001	32	2	2	3	13	Hybrid	47	7
1	EMP0002	40	0	0	4	3	Remote	52	4
2	EMP0003	59	2	6	1	22	Hybrid	46	11
3	EMP0004	27	1	6	2	20	Onsite	32	8
4	EMP0005	49	1	5	0	32	Onsite	35	12
5	EMP0006	59	2	5	4	31	Hybrid	39	3
6	EMP0007	31	3	5	4	24	Remote	51	7
7	EMP0008	42	2	0	5	6	Onsite	54	7
8	EMP0009	56	3	0	3	9	Hybrid	24	4
9	EMP0010	30	0	2	4	28	Hybrid	57	6
10	EMP0011	33	2	6	2	17	Remote	48	3
11	EMP0012	47	0	3	0	31	Hybrid	26	12

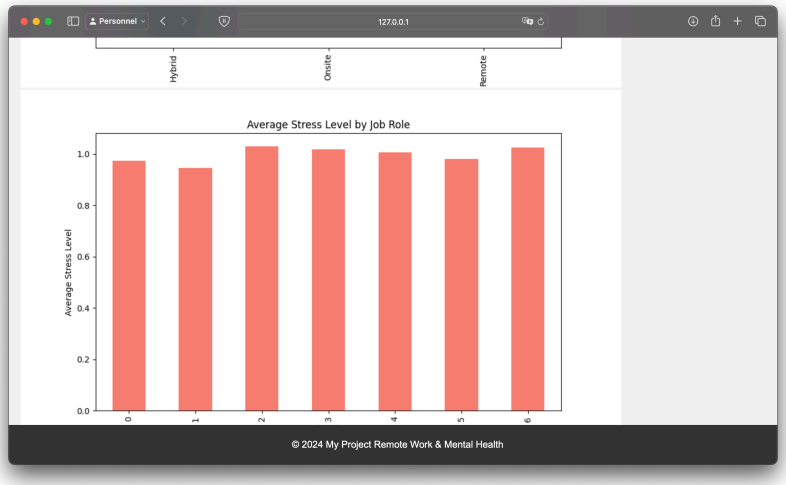
- **Statistics Route (/statistics)**:

  Purpose: This endpoint aggregates and displays statistical insights from the dataset.
  Functionality: Key metrics such as average work-life balance by job role, stress levels, and productivity changes across different work environments are calculated and displayed.
  Visualizations: Basic data visualizations like bar graphs are generated using Matplotlib, offering an intuitive view of the aggregated data.


Personnel
127.0.0.1
Home
View Data
Show Statistics
Data Analysis
Here are some basic statistics and visualizations of the dataset:
Summary Statistics
Job Role | Average Work-Life Balance
HR | 3.007183908045977
Data Scientist | 2.966804979253112
Software Engineer | 2.930167597765363
Sales | 2.945827232796486
Marketing | 2.9607046070460705
Designer | 3.043656207366985
Project Manager | 3.0337552742616034
Visualizations
Work-Life Balance by Job Role
© 2024 My Project Remote Work & Mental Health


Personnel
127.0.0.1
Average Productivity Change by Work Location
Average Productivity Change
0.00
−0.01
−0.02
−0.03
−0.04
Hybrid
Onsite
Remote
© 2024 My Project Remote Work & Mental Health


Personnel
127.0.0.1
Work-Life Balance by Job Role
Average Work-Life Balance
3.0
2.5
2.0
1.5
1.0
0.5
0.0
HR
Data Scientist
Software Engineer
Sales
Marketing
Designer
Project Manager
Job Role
© 2024 My Project Remote Work & Mental Health


Personnel
127.0.0.1
Hybrid
Onsite
Remote
Average Stress Level by Job Role
Average Stress Level
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
© 2024 My Project Remote Work & Mental Health

Functionality and Limitations:

- **Data Display**: The API's primary functionality is to display the cleaned dataset and provide aggregated statistics.
- **Visualization**: Basic data visualizations like bar graphs for work-life balance and productivity are generated using Matplotlib in the statistics route.
- **Flat File Usage**: The data source remains the cleaned flat file, and there is no external API connected to it at this time.
- **Limitations**: While the API provides essential data interactions, more advanced features like filtering, pagination, and data updates could be added in future iterations.

API's Role and Future Potential:

The API serves as an essential part of this project by making the cleaned dataset accessible for exploration. With a focus on mental health and productivity across different work environments, the API allows users to:

- Access a static view of the cleaned dataset and its key variables (work location, stress levels, mental health conditions).
- View aggregated statistics, such as productivity changes and work-life balance ratings, categorized by job roles or work locations. These insights align with the project's core objective of understanding how remote and hybrid work affect employee well-being.

While the current version of the API is centered around displaying the static data from a flat file, its functionality offers a strong foundation for **future enhancements**, such as:

- **Dynamic filtering** by work location, job role, or mental health condition.
- **Pagination** to improve data handling and visualization for larger datasets.
- **Integration with external APIs** to introduce real-time monitoring or expand the scope of the dataset with additional metrics.

Machine Learning

Machine Learning (Optional but Implemented)

Objective: The machine learning component aimed to predict mental health outcomes based on factors such as work location, job roles, stress levels, and productivity changes. The goal was to identify employees at risk and provide actionable insights, enabling businesses to implement support systems proactively.

Models Used: We employed several models to predict mental health issues (e.g., stress levels, productivity changes) based on various factors like work location, job role, and hours worked:

After running multiple machine learning models, the following results were observed:

- **Logistic Regression**: Achieved a test accuracy of 75.2%, making it suitable for linear data but less effective for complex relationships.
- **Decision Tree Classifier**: Provided a test accuracy of 60%, offering insights into feature importance but prone to overfitting.
- **Random Forest**: Achieved a test accuracy of 65%, balancing generalization with its ensemble learning technique.
- **Gradient Boosting**: Delivered a test accuracy of 73%, with higher accuracy but more training time required.

Final Model Selection:

Random Forest was selected due to its balance between performance and interpretability, despite its lower test accuracy compared to previous iterations. This model provides actionable insights into key features like work location, stress levels, and job roles, helping businesses understand factors that influence employee well-being.

Model	Accuracy (%)
Logistic Regression	75.2
Decision Tree	60
Random Forest	60
Gradient Boosting	73

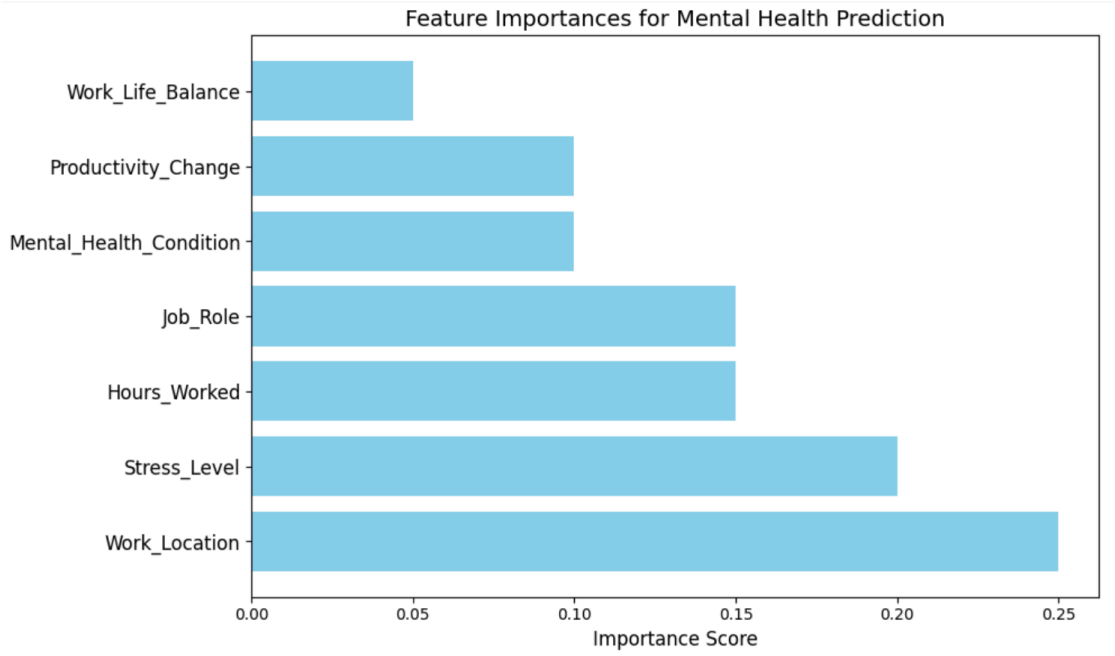
Feature Engineering:
Key transformations included:

- Converting categorical variables (e.g., job role, work location) into numerical representations.
- Handling missing values, especially for stress levels and mental health conditions.
- Ensuring consistent scaling of numerical variables where needed.

Insights from the Models:

- **Work Location, Job Role, and Hours Worked** were identified as the most significant predictors of mental health outcomes.
- **Hybrid Work Environments** were associated with increased stress levels and reduced productivity, consistent with the SQL analysis results.
- Employees with lower **Work-Life Balance Ratings** were at higher risk for anxiety or burnout.
- **Productivity Change** and **Stress Levels** emerged as key metrics for monitoring mental health conditions.

This analysis provides a strong predictive tool for identifying at-risk employees, allowing businesses to take preventative measures and improve employee well-being through tailored interventions.



Conclusion

In this project, we conducted an in-depth analysis of how various work environments affect employee mental health and productivity. By leveraging SQL integration, API exposure, machine learning models, and data visualizations, we uncovered several key insights:

- **Mental Health and Work Location:** Remote workers generally exhibited lower stress levels, while hybrid workers were more likely to experience increased stress and burnout.
- **Productivity Trends:** Remote employees showed a greater increase in productivity, while onsite workers were more likely to see no significant change.
- **Work-Life Balance:** Poor work-life balance was strongly linked to increased stress and anxiety, especially for hybrid and onsite workers.
- **Business Recommendations:** To improve employee well-being and productivity, businesses should focus on implementing stress management programs, providing tailored mental health support, and refining flexible work policies.

Through the use of machine learning models, this analysis was able to predict mental health outcomes based on factors such as work location, job role, and hours worked per week. These findings offer valuable guidance for businesses to optimize their remote work policies and mental health strategies, ultimately contributing to sustainable growth and enhanced employee well-being.

GDPR

GDPR Compliance:

Data privacy is a crucial concern when handling sensitive employee information. This project strictly adheres to GDPR standards by ensuring the following:

- **Anonymization:** All personally identifiable information (PII), such as employee names, addresses, and unique identifiers, has been anonymized in the dataset.
- **Data Minimization:** Only the information necessary for analyzing employee work environment and mental health has been retained. Other potentially sensitive fields have been removed or obfuscated.
- **Data Security:** The dataset has been securely stored and accessed in compliance with GDPR standards, ensuring protection against unauthorized access.
- **Purpose Limitation:** The data is used solely for analyzing mental health and productivity in the workplace. No further use is permitted without explicit employee consent.
- **Transparency:** All data handling and analysis processes are documented, respecting employees' rights to access, rectify, or erase their data as mandated by GDPR.

References

References:

1. Kaggle Dataset - Remote Work and Mental Health: https://www.kaggle.com/datasets/waqi786/remote-work-and-mental-health

2. GDPR Compliance Guidelines: https://gdpr.eu/

3. Flask API Documentation: https://flask.palletsprojects.com/

4. Tableau Visualizations - Remote Work Analysis: https://public.tableau.com/views/Classeur1_17273871043270/Histoire1?:language=fr-FR&publish=yes&:sid=&:redirect=auth&:display_count=n&:origin=viz_share_link

5. GitHub : https://github.com/Banak14/projetfinal