UNIVERSITY OF PORTSMOUTH

PROGRAMMING FOR DATA ANALYTICS AND AI

COURSEWORK

UP2225522

1. Introduction

The dataset examines the psychological impact of the COVID-19 pandemic on individuals across various demographics. It explores factors such as age, gender, occupation, work routines before and during the pandemic, travel time, adaptation to remote work, home environment satisfaction, productivity changes, sleep patterns, skill acquisition, family connections, relaxation levels, personal time, preferences for working from home or the office, and the desire for hybrid work models. These insights shed light on how different aspects of life were affected and adapted during this unprecedented period.

2. Data Preparation

2.1 Data Preparation for Descriptive Analytics

Table 1

	age	gender	occupation	line_of_work	time_bp	time_dp	travel_time	easeof_online	home_env	prod_inc	sleep_bal	new_skill	fam_connect	relaxed	relaxed	self_time	like_hw	dislike_hw	prefer	certaindays_hw
0	19-25	Male	Student in College	NaN	7	5	0.5	3	3	0.0	0.0	0.5	1.0	-0.5	-0.5	-0.5	100	1	Complete Physical Attendance	Yes
1	Dec-18	Male	Student in School	NaN	7	11	0.5	4	2	-0.5	0.5	-1.0	1.0	1.0	1.0	1.0	1111	1110	Complete Physical Attendance	No
2	19-25	Male	Student in College	NaN	7	7	1.5	2	2	1.0	0.0	0.5	0.5	0.5	0.5	0.5	1100	111	Complete Physical Attendance	Yes
3	19-25	Male	Student in College	NaN	7	7	1.5	3	1	0.0	1.0	0.5	0.0	-1.0	-1.0	-0.5	100	1111	Complete Physical Attendance	Yes
4	19-25	Female	Student in College	NaN	7	7	1.5	2	2	0.0	0.0	0.0	0.0	0.5	0.5	0.0	1010	1000	Complete Physical Attendance	Yes

This table shows the top five rows of the data. It can be seen that there are null values and also some data which will need to be replaced.

Table 2

```
RangeIndex: 1175 entries, 0 to 1174
Data columns (total 19 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   age              1175 non-null   object
 1   gender           1175 non-null   object
 2   occupation       1175 non-null   object
 3   line_of_work     479 non-null    object
 4   time_bp          1175 non-null   int64
 5   time_dp          1175 non-null   int64
 6   travel_time      1175 non-null   float64
 7   easeof_online    1175 non-null   int64
 8   home_env         1175 non-null   int64
 9   prod_inc         1175 non-null   float64
 10  sleep_bal        1175 non-null   float64
 11  new_skill        1175 non-null   float64
 12  fam_connect      1175 non-null   float64
 13  relaxed          1175 non-null   float64
 14  self_time        1175 non-null   float64
 15  like_hw          1175 non-null   int64
 16  dislike_hw       1175 non-null   int64
 17  prefer           1175 non-null   object
 18  certaindays_hw   1175 non-null   object
dtypes: float64(7), int64(6), object(6)
memory usage: 174.5+ KB
```

```
psyco1.shape
✓ 0.0s
(1175, 19)
```

Based on Table 2, there are 19 columns with 1175 rows and column ‘line_of_work’ contains 479 non-null rows, meaning there are 696 null rows.

Figure 1

```
psyco1[psyco1['occupation'] != 'Working Professional']['line_of_work'].isnull().sum()
✓ 0.0s
696
```

After further investigation, it is shown that all the null rows are people with occupations as not working professionals as shown in Figure 1. These null values are replaced with ‘other/not working’.

Table 3

```
19-25     345
26-32     261
40-50     181
50-60     170
33-40     102
Dec-18     74
60+        42
Name: age, dtype: int64
```

Column ‘age ‘ is expanded further to look at the value counts and is shown that one of the age ranges is labelled as ‘Dec-18’. These rows are replaced with ‘0 – 18’.

Tables 4 and 5 reflect these changes.

Table 4

```
psyco1.isnull().sum()
✓ 0.0s
age               0
gender            0
occupation        0
line_of_work      0
time_bp           0
time_dp           0
travel_time       0
easeof_online     0
home_env          0
prod_inc          0
sleep_bal         0
new_skill         0
fam_connect       0
relaxed           0
self_time         0
like_hw           0
dislike_hw        0
prefer            0
certaindays_hw    0
dtype: int64
```

2.2 Data Preparation for Classification

Based on Table 6, the amount of people who input their gender as 'Prefer not to say' is very low (8 in total, less than 1% of the data). These results are half replaced as 'Male' and the other half as 'Female'. The table below reflects the change.

Table 5

```
psyco2['gender'].value_counts()
✓ 0.0s

Male      653
Female    522
Name: gender, dtype: int64
```

Classification is done on 'genders' with each value encoded as:

- Female: 0
- Male: 1

After several trials data preparation is done with different combinations to achieve the best results. The final results ended up using one hot encoding with these columns:

```
columns_to_encode = ['dislike_hw', 'certaindays_hw', 'occupation', 'line_of_work', 'prefer', 'like_hw']
```

And label encoding for the rest. The figure below shows the final results after data preparation.

Figure 2

	age	gender	time_bp	time_dp	travel_time	easeof_online	home_env	prod_inc	sleep_bal	new_skill	...	like_hw_5	like_hw_6	like_hw_7	like_hw_8	like_hw
0	1	1	2	1	0	2	2	2	2	3	...	0	0	0	0	
1	0	1	2	4	0	3	1	1	3	0	...	0	0	0	0	
2	1	1	2	2	1	1	1	4	2	3	...	0	0	0	0	
3	1	1	2	2	1	2	0	2	4	3	...	0	0	0	0	
4	1	0	2	2	1	1	1	2	2	2	...	0	0	0	0	

5 rows × 65 columns

Cross-validation is used to ensure that the model performs well on new, unseen data. By splitting the dataset into multiple parts, called folds, and training the model on different combinations of these folds, cross-validation helps assess how well the model generalizes to different data points. This technique gives a more reliable estimate of the model's performance and helps in detecting potential issues like overfitting, where the model becomes too specialized to the training data and doesn't perform well on new examples.

2.3 Data Preparation for Regression

Regression is done to predict 'self_time' and data preparation is also implemented with different combinations to achieve the best results. The final result is to use one hot encoding on 'prefer' and 'gender' and the rest is labelled encoded and then standardised. The figure below shows the final results after data preparation.

Figure 3

	age	occupation	line_of_work	time_bp	time_dp	travel_time	easeof_online	home_env	prod_inc	sleep_bal	...	relaxed	self_time	like_hw
0	-0.924820	0.043083	0.316183	-0.20719	-1.118996	-0.740045	0.368080	0.200490	-0.014535	0.175435	...	-0.855317	-1.077189	-1.357072
1	-1.520578	0.487135	0.316183	-0.20719	1.140145	-0.740045	1.157303	-0.609048	-0.827779	0.980651	...	1.539434	1.694410	0.804100
2	-0.924820	0.043083	0.316183	-0.20719	-0.365949	0.662459	-0.421143	-0.609048	1.611954	0.175435	...	0.741184	0.770544	0.780585
3	-0.924820	0.043083	0.316183	-0.20719	-0.365949	0.662459	0.368080	-1.418586	-0.014535	1.785868	...	-1.653567	-1.077189	-1.357072
4	-0.924820	0.043083	0.316183	-0.20719	-0.365949	0.662459	-0.421143	-0.609048	-0.014535	0.175435	...	0.741184	-0.153322	0.588196

5 rows × 22 columns

2.4 Data preparation for Clustering

Data preparation is done by label encoding all the text value columns to numbers and then standardising all columns. The figure below shows the final result ready for clustering.

Figure 4

	age	gender	occupation	line_of_work	time_bp	time_dp	travel_time	easeof_online	home_env	prod_inc	sleep_bal	new_skill	fam_connect
0	-0.924820	0.852431	0.043083	0.316183	-0.20719	-1.118996	-0.740045	0.368080	0.200490	-0.014535	0.175435	0.548935	1.077260
1	-1.520578	0.852431	0.487135	0.316183	-0.20719	1.140145	-0.740045	1.157303	-0.609048	-0.827779	0.980651	-1.782386	1.077260
2	-0.924820	0.852431	0.043083	0.316183	-0.20719	-0.365949	0.662459	-0.421143	-0.609048	1.611954	0.175435	0.548935	0.348963
3	-0.924820	0.852431	0.043083	0.316183	-0.20719	-0.365949	0.662459	0.368080	-1.418586	-0.014535	1.785868	0.548935	-0.379334
4	-0.924820	-1.111503	0.043083	0.316183	-0.20719	-0.365949	0.662459	-0.421143	-0.609048	-0.014535	0.175435	-0.228172	-0.379334

3. Descriptive Analytics

Table 6

```
Column: age
Number of unique values: 7
Unique values:
19-25    345
26-32    261
40-50    181
50-60    170
33-40    102
0-18      74
60+       42
Name: age, dtype: int64
```

The dataset covers seven age groups, with the highest representation in the 19-25 category, followed by 26-32, and then 40-50. There's a diverse spread across different age brackets, including individuals from younger groups (0-18) and older ones (60+).

Table 7

```
Column: gender
Number of unique values: 3
Unique values:
Male                649
Female              518
Prefer not to say     8
Name: gender, dtype: int64
```

Three gender categories are represented, with a majority of respondents identifying as Male, followed by Female. There's a small subset of respondents who preferred not to disclose their gender.

Table 8

```
Column: occupation
Number of unique values: 8
Unique values:
Working Professional      479
Student in College        358
Entrepreneur              119
Homemaker                  82
Medical Professional       73
Currently Out of Work      44
Student in School          18
Retired/Senior Citizen      2
Name: occupation, dtype: int64
```

Eight different occupations are included, with Working Professionals being the most prominent group, followed by Students in College. The dataset also has Entrepreneurs, Homemakers, Medical Professionals, and others.

Table 9

```
Column: line_of_work
Number of unique values: 9
Unique values:
Other/not working      696
Teaching               217
Engineering            116
Management              66
Other                   40
Government Employee     35
Architect                3
APSPDCL                  1
Architecture             1
Name: line_of_work, dtype: int64
```

Nine distinct categories of work roles are represented. The largest category is 'Other/Not Working,' followed by Teaching and Engineering. It covers a variety of fields such as Management, Government Employment, and specialized roles like Architecture and Electrical Services (APSPDCL).

Table 10

```
Column: prefer
Number of unique values: 2
Unique values:
Complete Physical Attendance    836
Work/study from home            339
Name: prefer, dtype: int64
```

There's a clear preference for complete physical attendance at work/study over working remotely, but a significant number of respondents favour a work/study from home arrangement.

Table 11

```
Column: certaindays_hw
Number of unique values: 3
Unique values:
Yes      568
No       309
Maybe    298
Name: certaindays_hw, dtype: int64
```

Respondents are divided on the necessity of specific days for working from home, with 'Yes,' 'No,' and 'Maybe' responses indicating varied preferences.

Table 12

```
like_hw  dislike_hw
100      1111          73
1000     1000          70
1110     101           52
1100     1111          50
110      1             47
                       ..
1000     111            1
10       1001           1
101      1010           1
10       11             1
11       1011           1
Length: 123, dtype: int64
```

The table above shows that the columns values for “like_hw” and “dislike_hw” does not make much sense. Harikrishnan, H (2023) the data creator suggest explains that there was a problem with how the information was encoded in the form. A higher value in the "like" binary, means a higher value in reality. It's recommended to consider using methods like min-max scaling to handle this. Tables 12 and 13 reflect these changes.

Table 13

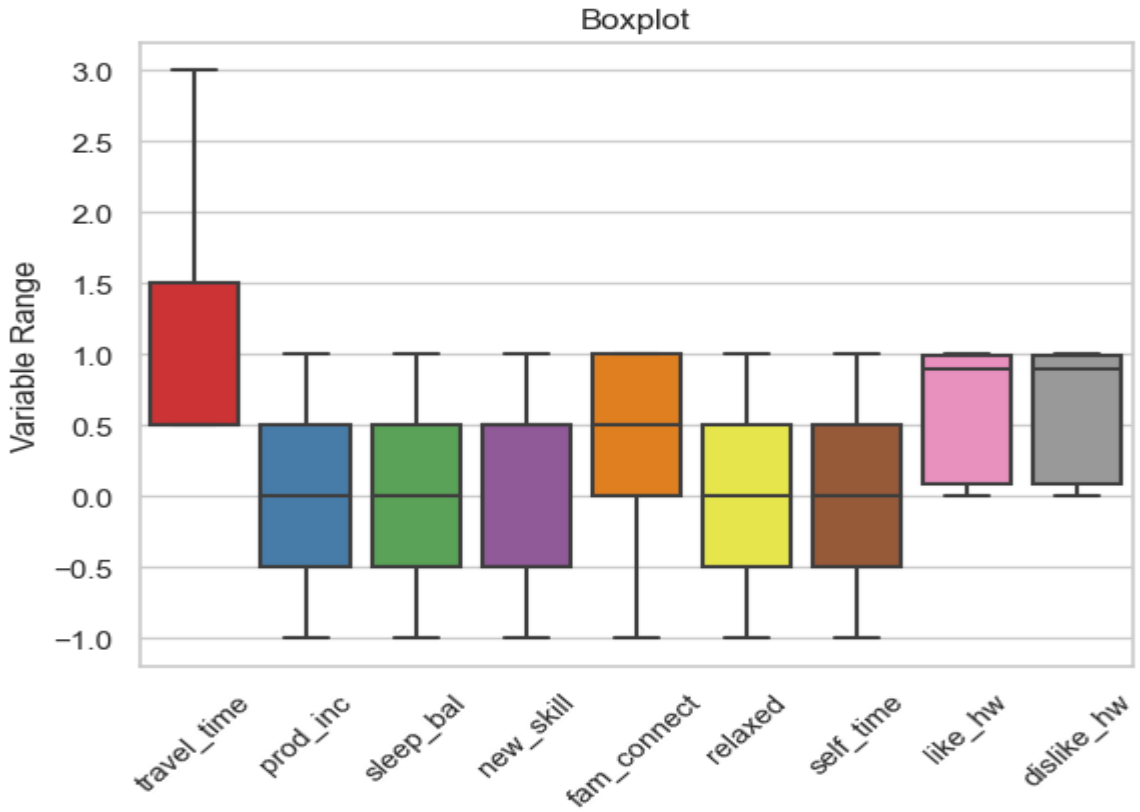
like_hw	dislike_hw	
0.089189	1.000000	73
0.900000	0.900000	70
0.999099	0.090090	52
0.990090	1.000000	50
0.098198	0.000000	47
		..
0.900000	0.099099	1
0.008108	0.900901	1
0.090090	0.909009	1
0.008108	0.009009	1
0.009009	0.909910	1

Length: 123, dtype: int64

Table 14

	time_bp	time_dp	travel_time	easeof_online	home_env	prod_inc	sleep_bal	new_skill	fam_connect	relaxed	self_time	like_hw	dislike_hw
count	1175.000000	1175.000000	1175.000000	1175.000000	1175.000000	1175.000000	1175.000000	1175.000000	1175.000000	1175.000000	1175.000000	1175.000000	1175.000000
mean	7.415319	7.971915	1.027660	2.533617	2.752340	0.008936	-0.108936	0.146809	0.260426	0.035745	0.082979	0.661118	0.585646
std	2.005385	2.657007	0.713314	1.267609	1.235799	0.615083	0.621215	0.643686	0.686825	0.626637	0.541434	0.421622	0.452540
min	4.000000	4.000000	0.500000	1.000000	1.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000	0.000000	0.000000
25%	5.000000	5.000000	0.500000	1.000000	2.000000	-0.500000	-0.500000	-0.500000	0.000000	-0.500000	-0.500000	0.089189	0.090090
50%	7.000000	9.000000	0.500000	2.000000	3.000000	0.000000	0.000000	0.500000	0.500000	0.000000	0.000000	0.900901	0.900000
75%	9.000000	9.000000	1.500000	4.000000	4.000000	0.500000	0.500000	0.500000	1.000000	0.500000	0.500000	0.990090	0.990991
max	12.000000	12.000000	3.000000	5.000000	5.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
mode	7.000000	9.000000	0.500000	1.000000	3.000000	0.500000	-0.500000	0.500000	0.500000	0.000000	0.000000	0.089189	1.000000

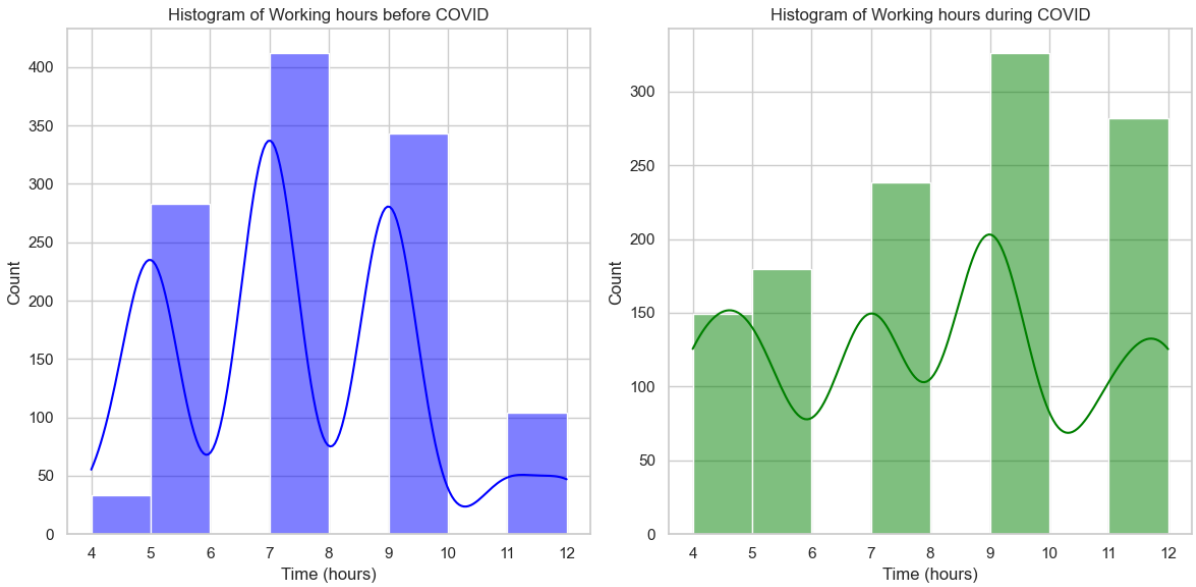
Figure 5



The boxplots and statistical summary table above shows various variable ranges for each numerical columns. For 'travel_time,' the majority spent around 0.5 to 1.5 hours, with few outliers extending up to 3 hours. 'prod_inc' (Productivity Increase) spans evenly around 0, indicating a mixed experience with some reporting a decrease, and others an increase. 'sleep_bal' shows a spread around 0, denoting a balanced distribution between improved and reduced sleep patterns. 'new_skill' showcases varied experiences, with the middle half showing gains in acquiring new skills. 'fam_connect' fluctuates. However, most respondents reported some level of connection with their families. 'relaxed' seems evenly distributed around 0, suggesting a mixed sentiment regarding relaxation levels. In addition, 'self_time' shows a balanced distribution between respondents feeling they gained or lost personal time during the pandemic.

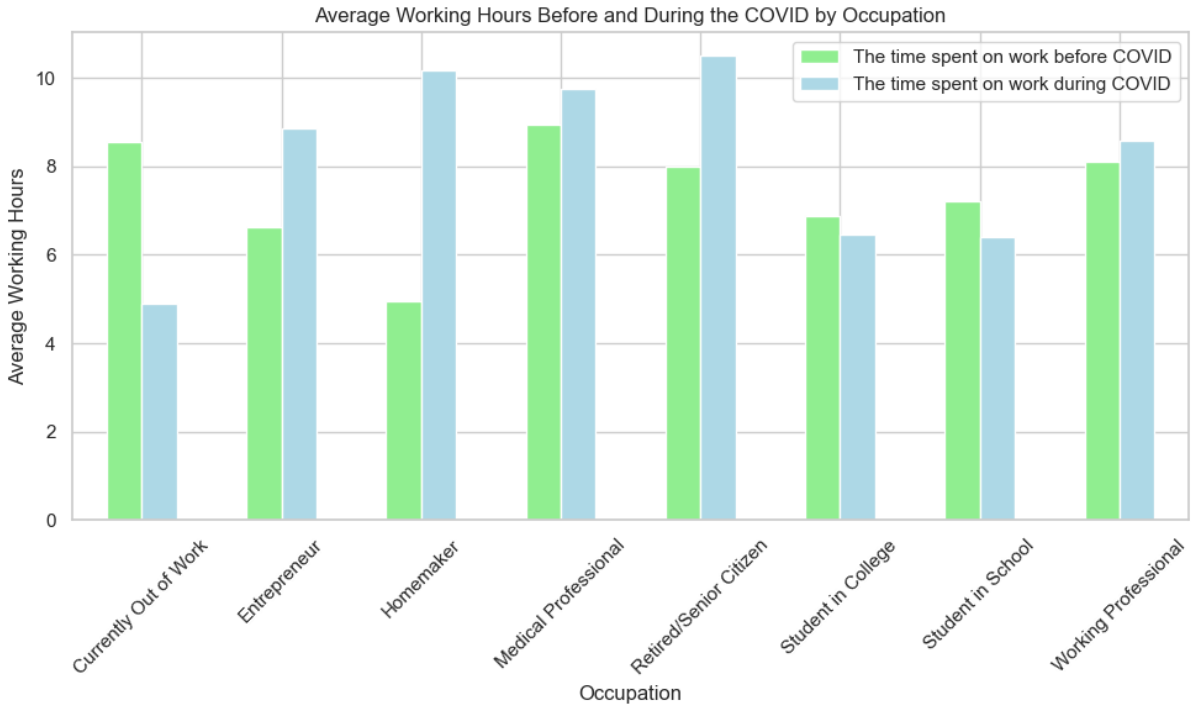
Lastly, focusing on the sentiments around working from home, the dataset shows a nuanced that; on average, individuals seemed to moderately favour working from home, with an average liking score of around 0.66, but this sentiment was supported by a moderate level of dislike, averaging around 0.59. The boxplot of 'like_hw' and 'dislike_hw' are negatively skewed, this indicates that most people either liked or didn't like working from home. Even though there might be a few who felt differently, most responses leaned towards extreme feelings. This tells us that people's opinions about working from home were often very strong, either positive or negative, with fewer people having a moderate view of it.

Figure 6



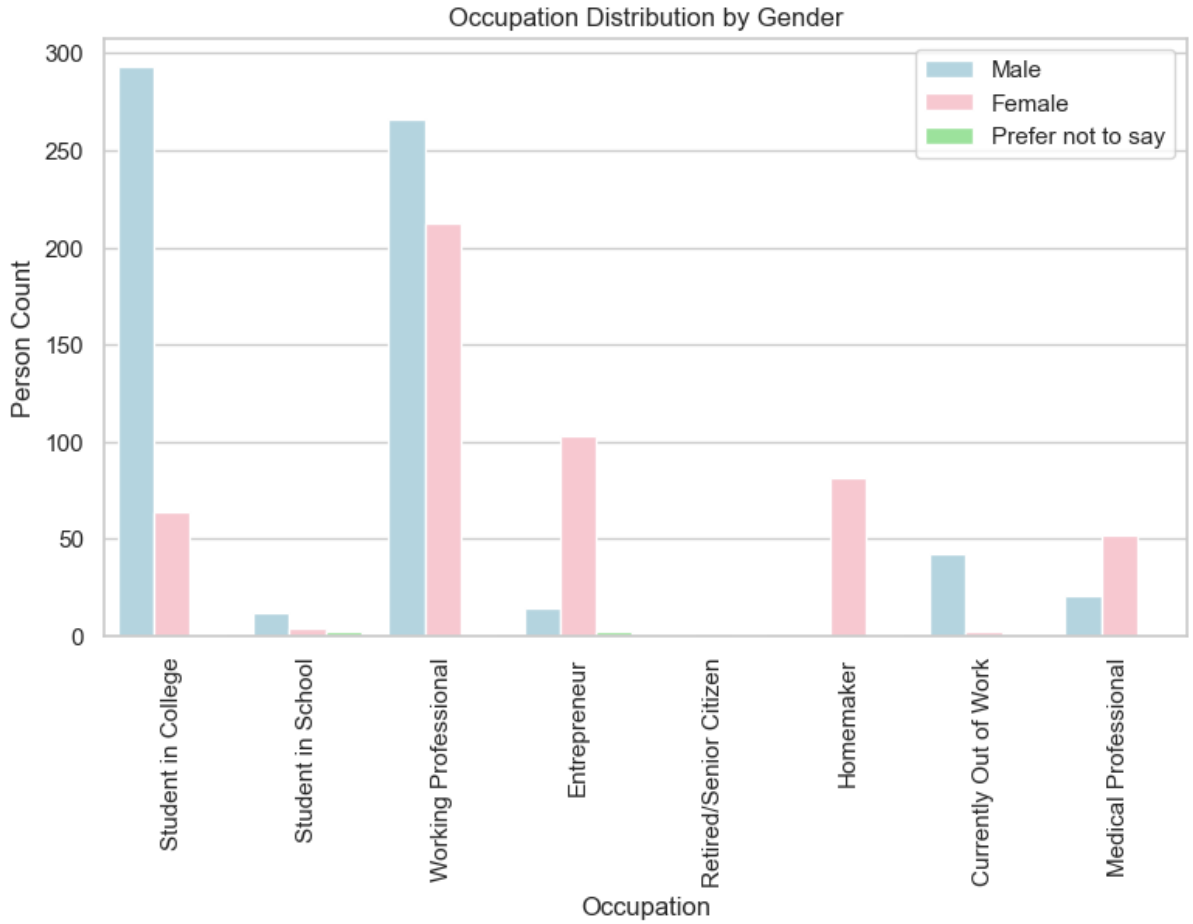
These histograms illustrate the distribution of time spent on work before and during the pandemic across six time brackets each. This comparison suggests that while the general distribution of work hours remained somewhat similar. However, there's a shift in the most common time during the pandemic, with a higher number of respondents working 9 hours compared to the other durations, unlike the pre-pandemic scenario where 7 hours was the most common.

Figure 7



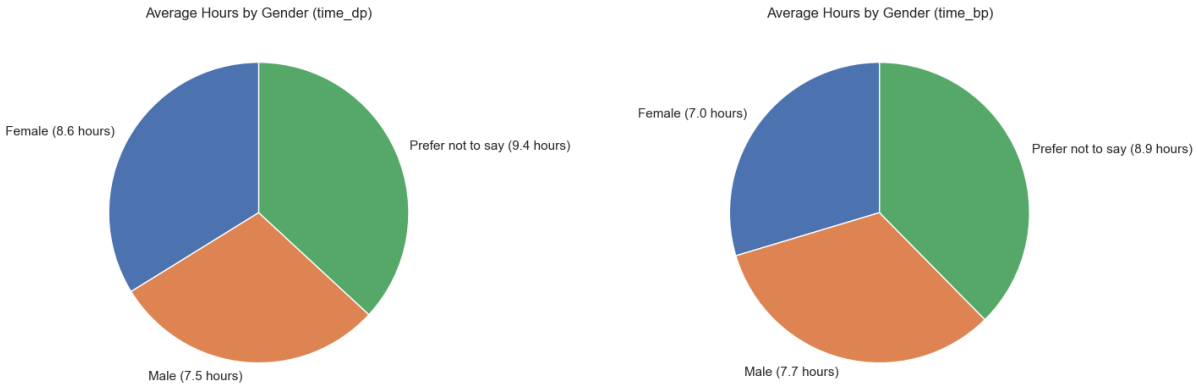
Across various occupations, the time spent on work before and during the pandemic reveals distinctive patterns. Homemakers, typically spending fewer hours on work initially, notably increased their work time during the pandemic, surpassing ten hours. Retired/Senior Citizens, who started with moderate work hours pre-pandemic, significantly increased their work time to around ten and a half hours during the pandemic. Similarly, Entrepreneurs, who initially had moderate work hours, notably increased their workload to almost nine hours during the pandemic. Medical Professionals, already spending substantial time on work before the pandemic, maintained high work hours during the pandemic, around nine and a half hours. On the other hand, Students in College and School slightly reduced their work hours during the pandemic, whereas Working Professionals maintained a consistent workload, showing only a slight increase during the pandemic. Interestingly, those currently out of work decreased their already lower work time further during the pandemic, showing a significant reduction compared to other occupational groups. These comparisons underline a wide range of adjustments in work hours across different occupations in response to the pandemic's challenges and changes in work dynamics.

Figure 8



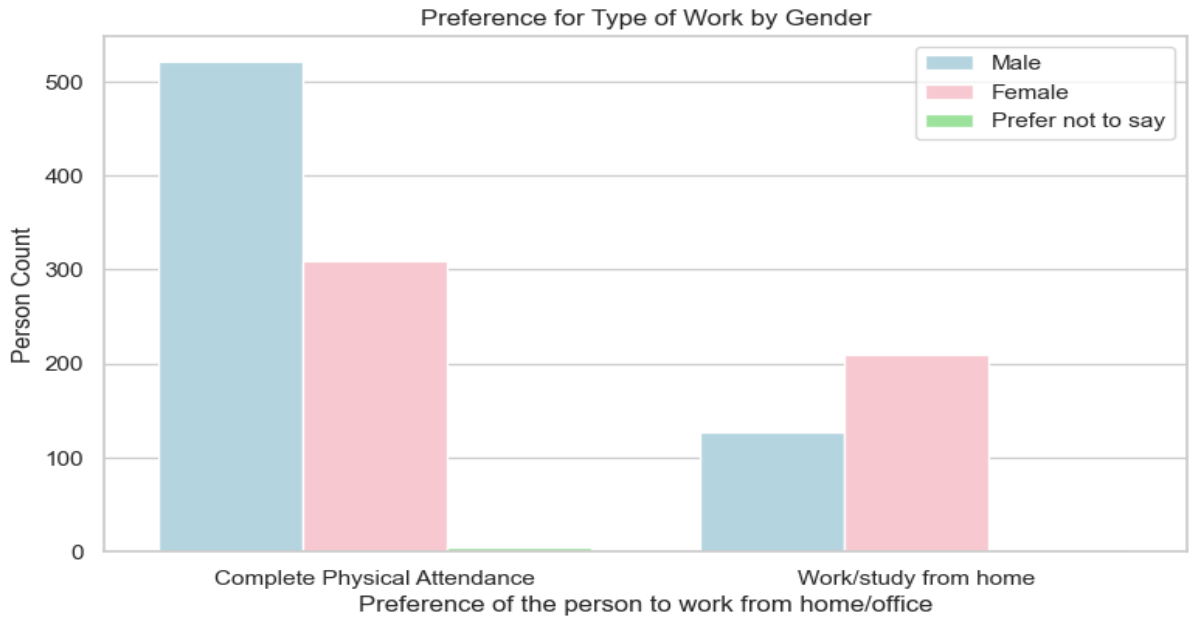
These figure above shows a comparisons which provide insights into the distribution of genders across different occupational categories, highlighting variations in participation levels between males and females across various roles. Among females, Working Professional roles had the highest count, followed by Entrepreneurship and Homemaking. Female representation was notable in Student roles, both in College and School, but comparatively lower. Only a few females were recorded as Currently Out of Work. On the other hand, among males, the count was highest for Students in College, closely followed by Working Professionals and those Currently Out of Work. Medical Professional roles had a moderate count for males, while Entrepreneurship and Student roles in School also had male representation, although relatively lower. Interestingly, Retired/Senior Citizen roles had minimal male representation, with just one male respondent in this category.

Figure 9



The average working hours, both during (time_dp) and before (time_bp) the pandemic, showcase interesting differences across gender categories and those who prefer not to disclose. During the pandemic, females worked the longest hours on average, clocking in at around 8.6 hours, while males worked slightly less, averaging about 7.5 hours. Individuals who preferred not to disclose their gender had the highest average working hours during the pandemic, recording an average of 9.4 hours. Before the pandemic, the trend shifted slightly, with females working around 7.0 hours on average, while males worked slightly more, averaging about 7.7 hours. Yet again, those who chose not to disclose their gender had the highest average working hours before the pandemic, at an average of 8.9 hours. These results, reflected in pie charts, would demonstrate a larger portion for females in average working hours during the pandemic, a slightly smaller portion for males, and a notably larger portion for those who preferred not to disclose their gender, emphasizing variations in average working hours based on gender identity and disclosure preferences both before and during the pandemic.

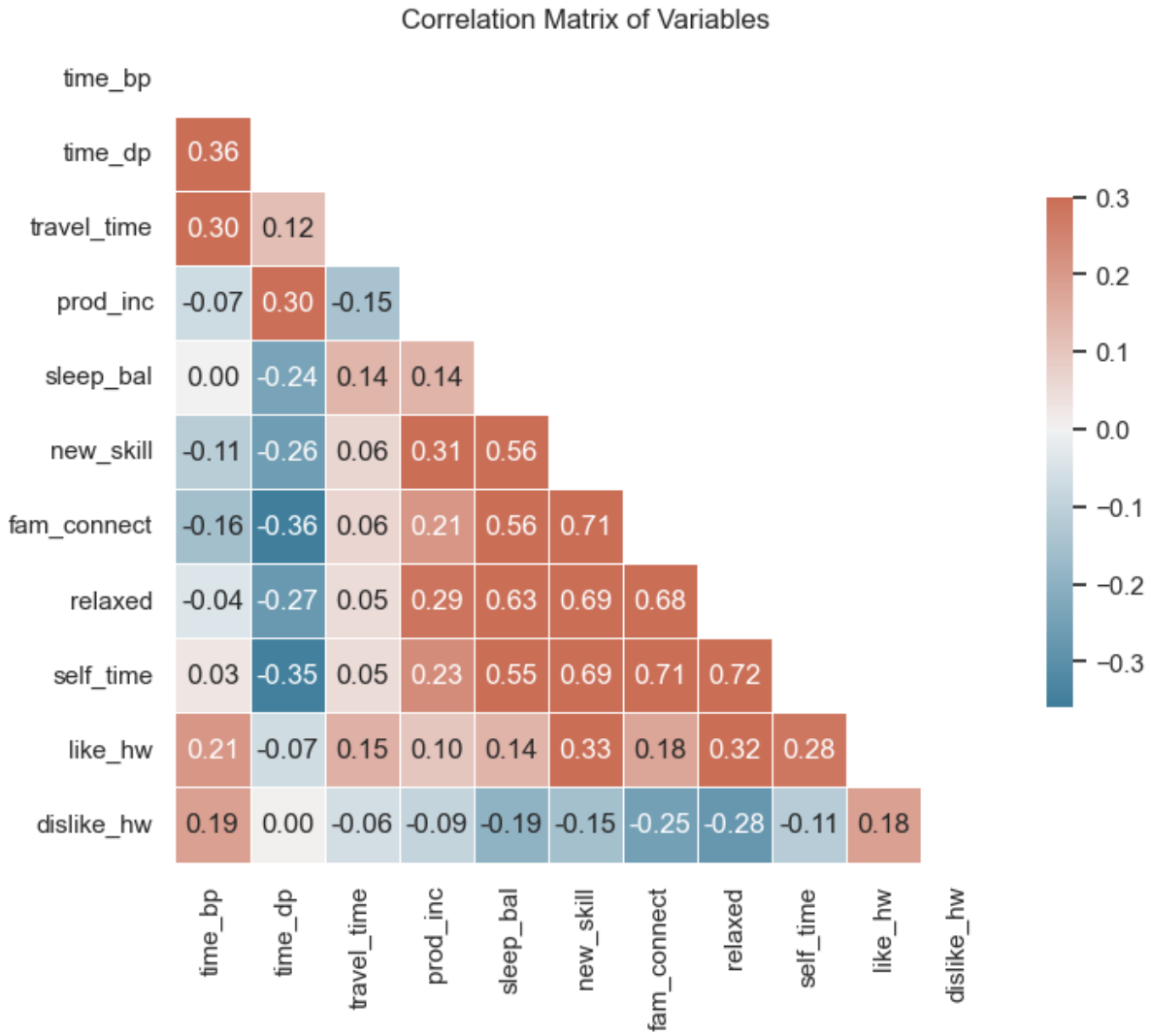
Figure 10



In the preferences for work or study arrangements, both males and females show a significant preference for attending physically at their workplace or educational institution. Among females,

there's a substantial count favouring Complete Physical Attendance, with a notable but comparatively smaller count preferring to work or study from home. Similarly, males most prefer Complete Physical Attendance, with a substantially higher count, while a smaller proportion opts for remote work or study. Interestingly, individuals choosing not to disclose their gender demonstrate a minimal presence in both preferences, with a very small count favouring either Complete Physical Attendance or remote work/study. These patterns underscore a clear preference for in-person attendance among both males and females, with a smaller yet existent preference for remote arrangements, while those choosing not to disclose their gender show notably fewer preferences in either direction.

Figure 11



- There's a moderate positive correlation between time spent before the pandemic (time_bp) and during the pandemic (time_dp), indicating that those who worked longer before the pandemic tended to continue longer hours during it. However, this correlation isn't very strong, suggesting other factors influencing changes in working hours.

- There's a mild positive correlation between travel time and various aspects, like productivity increase and dislike towards certain work aspects. This suggests that individuals with longer travel times might experience reduced productivity and more negative feelings about certain work elements.

- Variables like productivity increase, new skill acquisition, and family connection show positive correlations. This implies that an increase in productivity often coincides with acquiring new skills and having stronger family connections. Moreover, variables like relaxation and self-time positively correlate with each other, indicating that feeling more relaxed might align with having more personal time.

- Preferences for certain work aspects show mild correlations with other variables. For instance, liking certain aspects of work correlates with increased new skill acquisition, while disliking certain aspects tends to relate to decreased productivity and well-being.

Overall, the correlations give insights into how various factors interrelate within the dataset, highlighting potential connections between working hours, travel time, productivity, and personal well-being.

4. Classification

4.1 Decision Tree

Figure 12

Figure 13

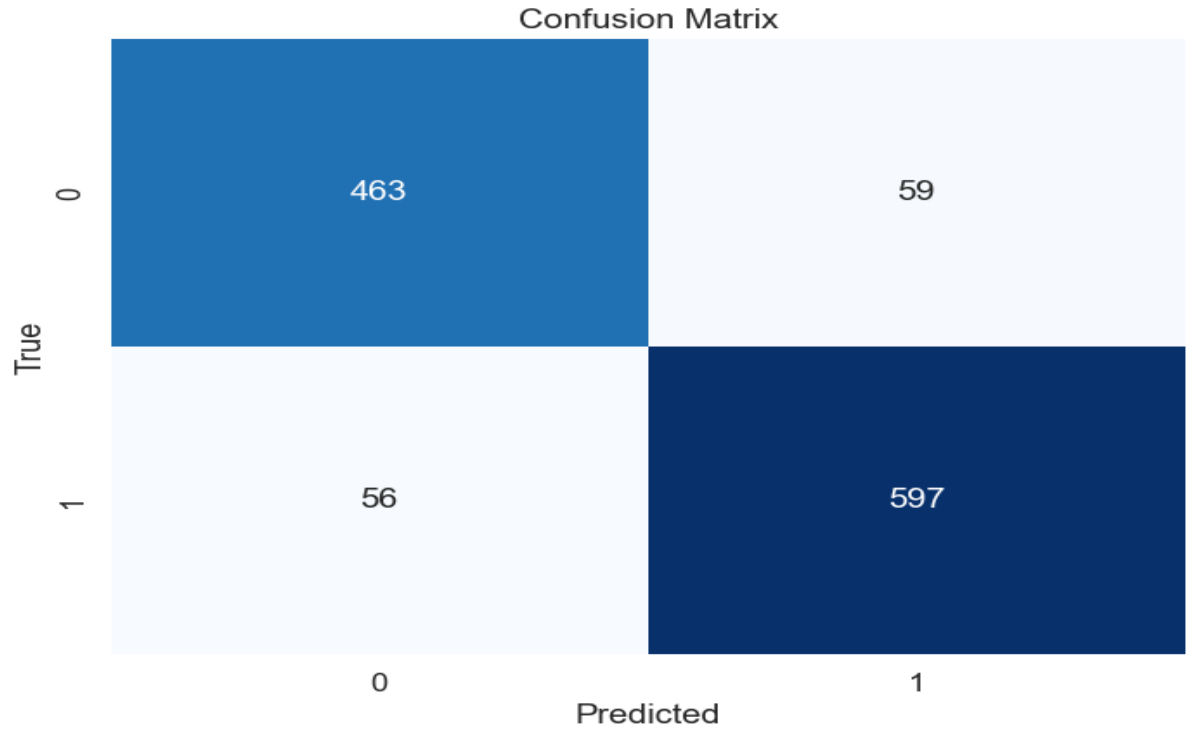


Table 15

```
Decision Tree - Cross-Validation AUC_score: 0.9006075701301979
Decision Tree - Cross-Validation Accuracy: 0.902127659574468
Decision Tree - Cross-Validation Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.89      0.89       522
           1       0.91      0.91      0.91       653

    accuracy                           0.90      1175
   macro avg       0.90      0.90      0.90      1175
weighted avg       0.90      0.90      0.90      1175

Decision Tree - Cross-Validation F1 Scores: [0.90091477 0.89076068 0.88829165 0.89720035 0.89284086]
Decision Tree - Cross-Validation Average F1 Score with variance: (0.8940016614097862, 2.0495684728893055e-05)
```

The decision tree model showcases a good overall performance in all metrics. It demonstrates balanced precision and recall for both genders, achieving F1-scores around 0.89 to 0.91. The variance in F1 scores is quite low, indicating consistency in performance. The type I &II error is distributed evenly suggesting a balance performance for this model.

4.2 KNN

Figure 14

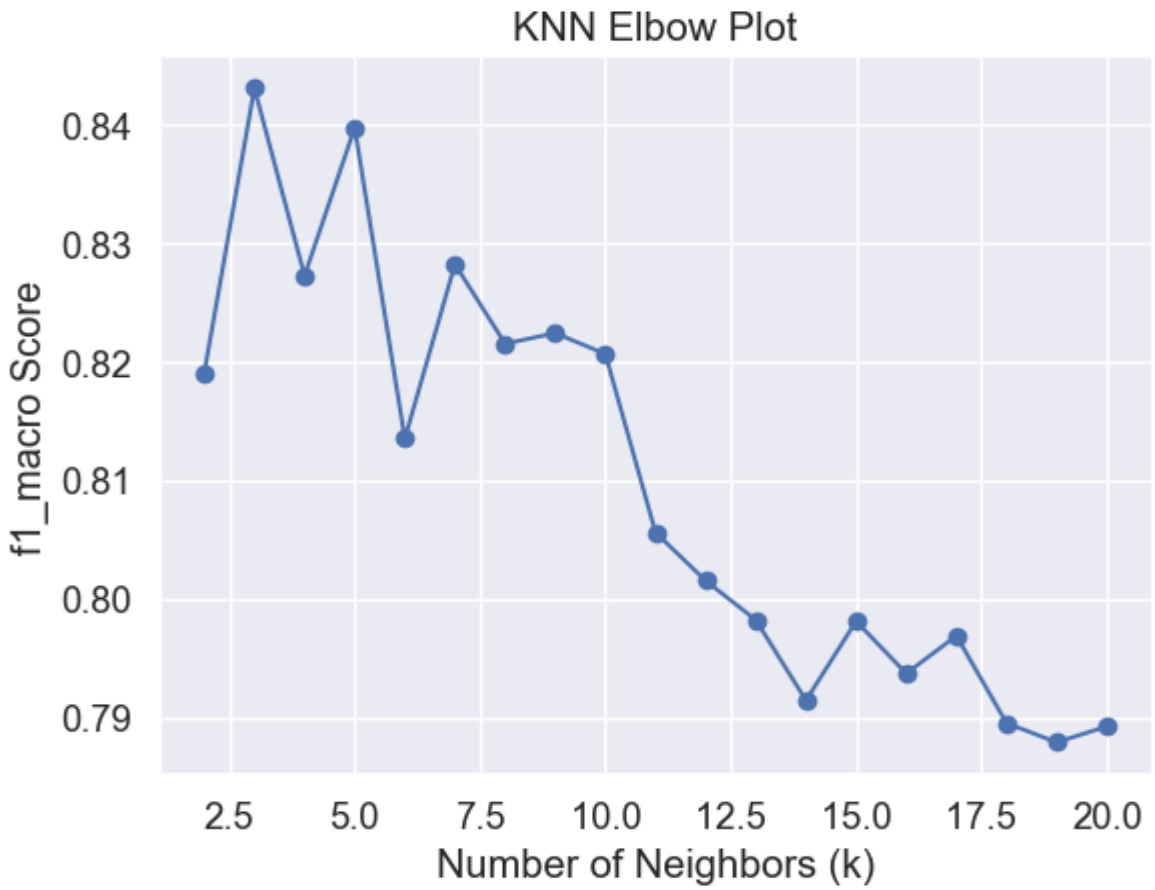


Figure 15

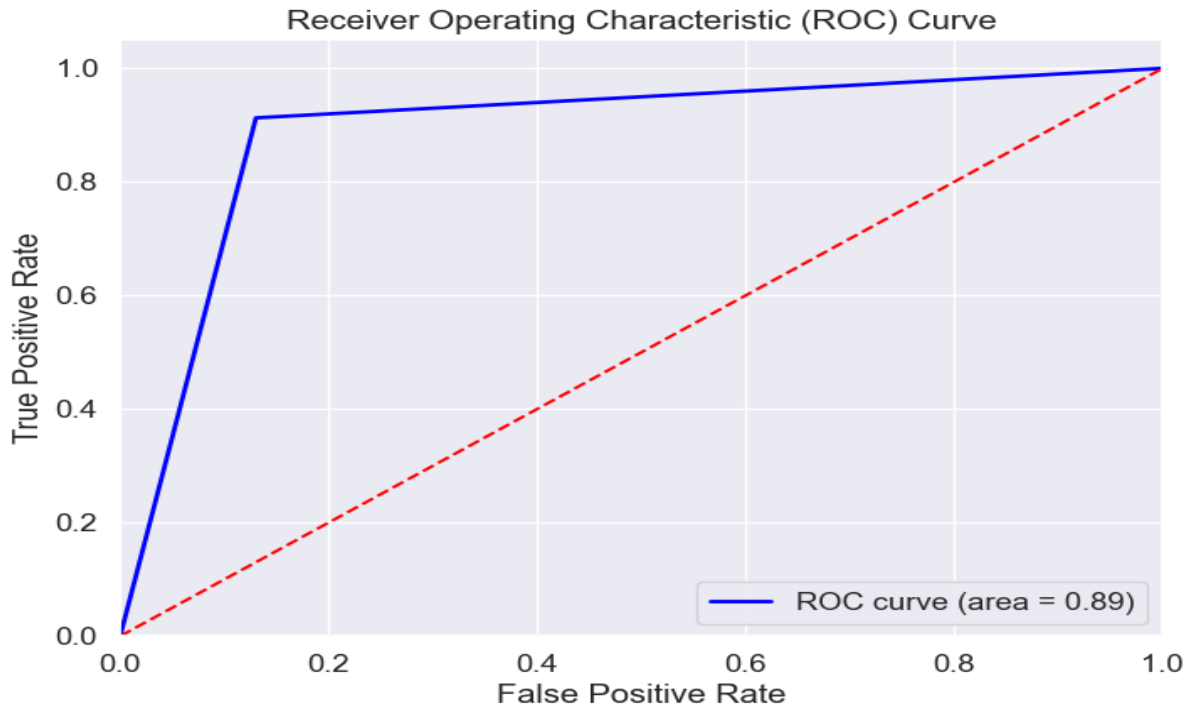


Figure 16

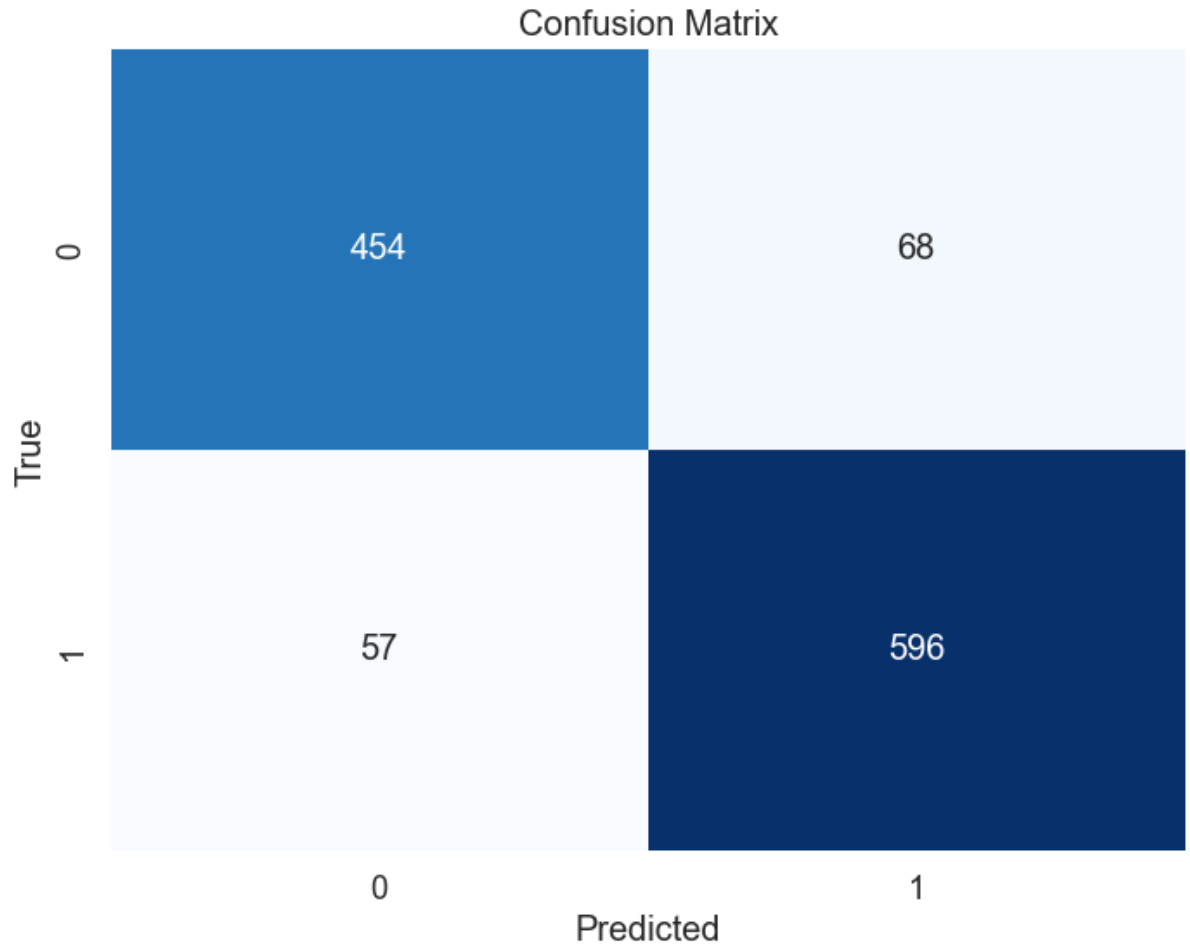


Table 16

```
Optimal k: 3
KNN - Cross-Validation AUC_score: 0.8912211836909518
KNN - Cross-Validation Accuracy: 0.8936170212765957
KNN - Cross-Validation Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.87      0.88       522
           1       0.90      0.91      0.91       653

    accuracy                           0.89      1175
   macro avg       0.89      0.89      0.89      1175
weighted avg       0.89      0.89      0.89      1175

KNN - Cross-Validation F1 Scores: [0.83557022 0.94015794 0.91323529 0.91436865 0.92169874 0.86299766
 0.93076923 0.87884615 0.88554444 0.83524791]
KNN - Cross-Validation Average F1 Score with variance: (0.8918436228185813, 0.001311625853457748)
```

The KNN model also performs well, with a good score in all metrics. Its precision and recall are relatively balanced, resulting in F1-scores ranging from 0.88 to 0.94. However, there's a bit more variance in its F1 scores compared to the decision tree. The type I &II error are nearly distributed evenly with a bias towards predicting Female.

4.3 Logistic Regression

Figure 17

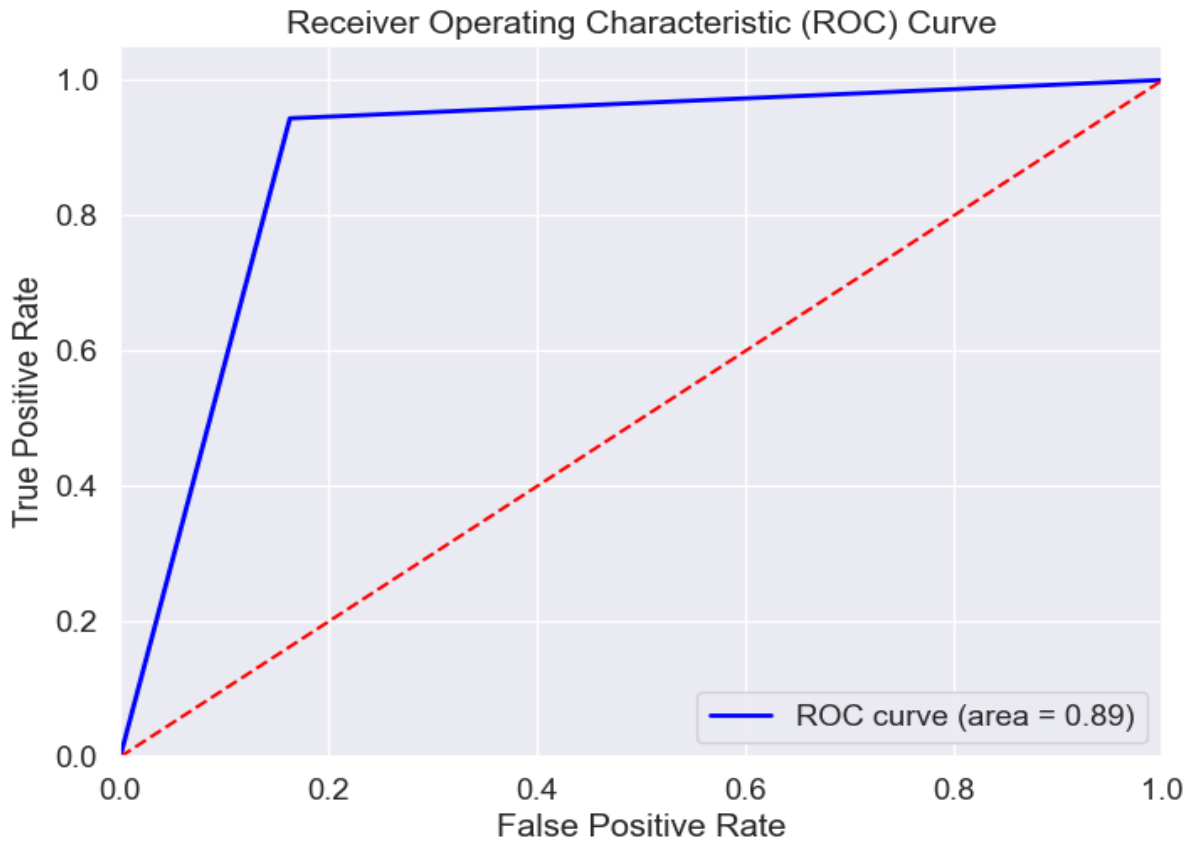


Figure 18

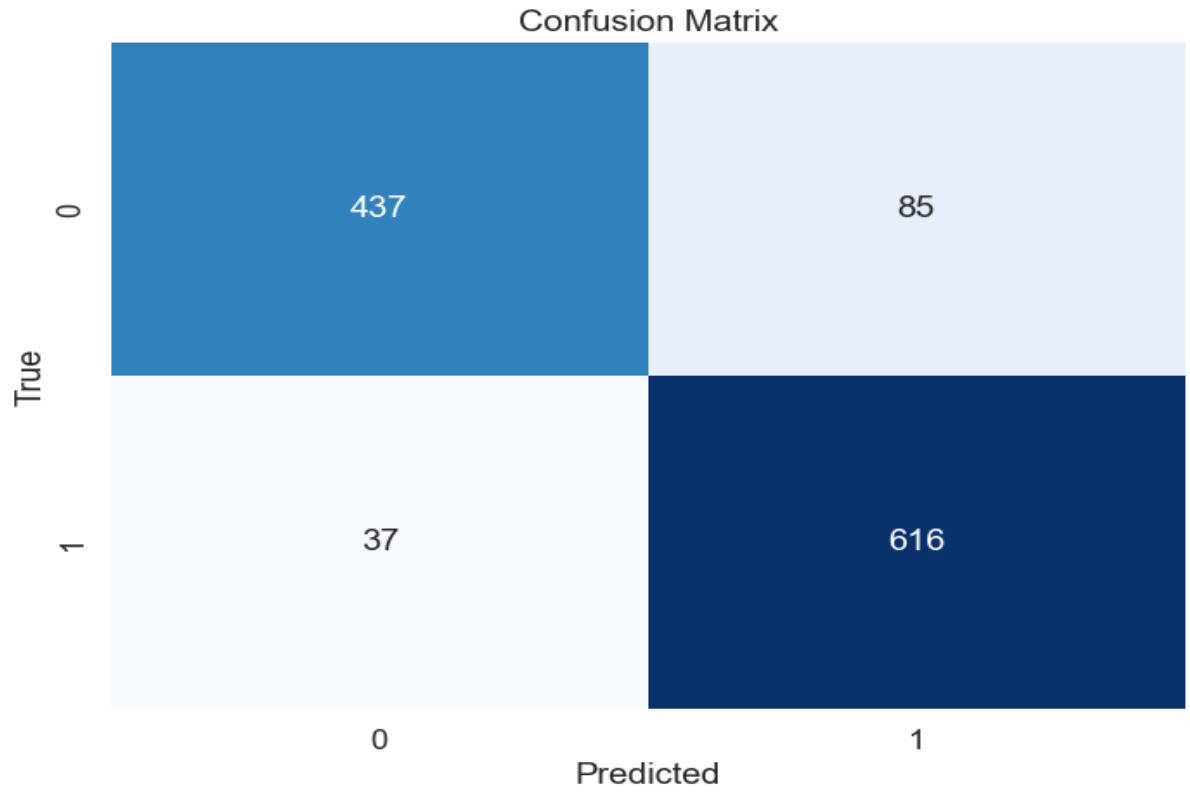


Table 17

```
Logistic - Cross-Validation AUC_score: 0.8902515944682075
Logistic - Cross-Validation Accuracy: 0.8961702127659574
Logistic - Cross-Validation Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.84      0.88       522
           1       0.88      0.94      0.91       653

    accuracy                           0.90      1175
   macro avg       0.90      0.89      0.89      1175
weighted avg       0.90      0.90      0.90      1175

Logistic - Cross-Validation F1 Scores: [0.85024263 0.94820018 0.91274771 0.89588235 0.90524856 0.86096257
 0.89402174 0.87834225 0.89465786 0.89522388]
Logistic - Cross-Validation Average F1 Score with variance: (0.8935529718154637, 0.0006670486563590575)
```

Logistic regression yields an AUC score of 0.89 and an accuracy of 0.90. It demonstrates a slightly lower precision for class 0 (female), indicating a bit more difficulty in correctly identifying females compared to males. Still, the F1-scores range between 0.88 to 0.91 with moderate variance. In addition, this model is bias towards classifying female based on the confusion matrix.

4.4 SVM

Figure 19

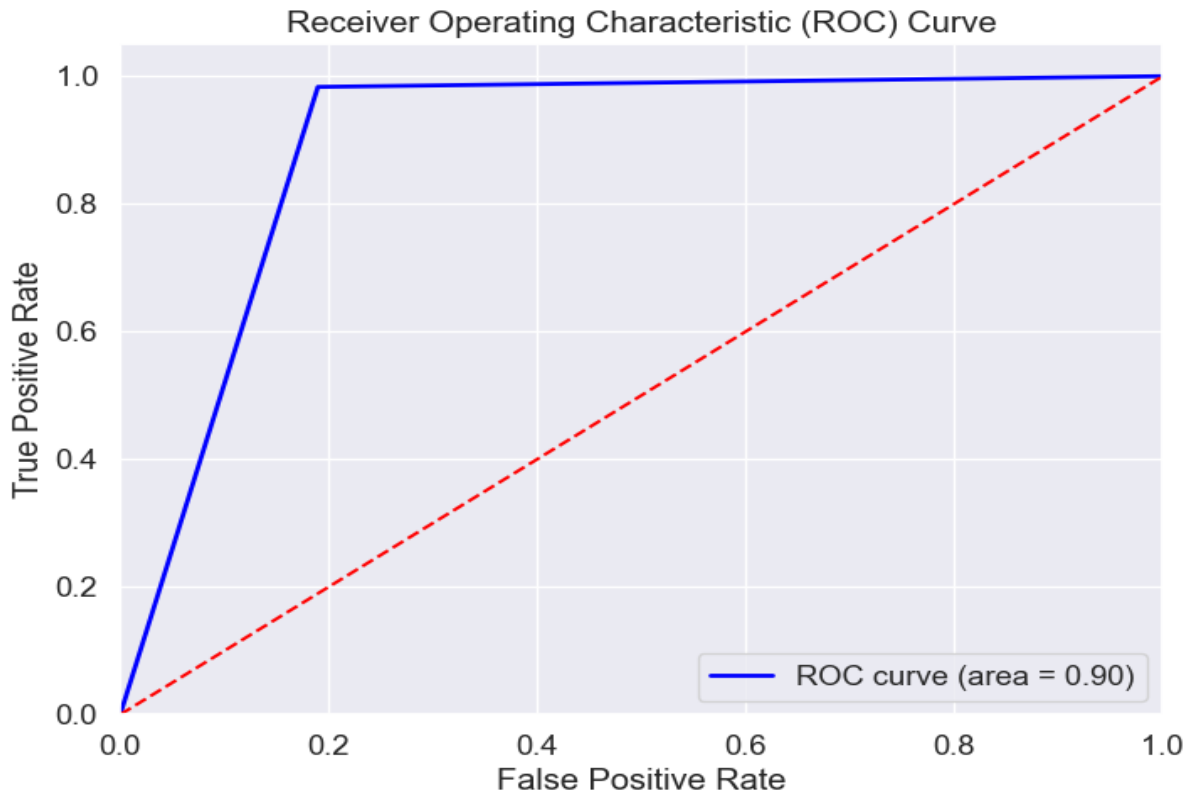


Figure 20

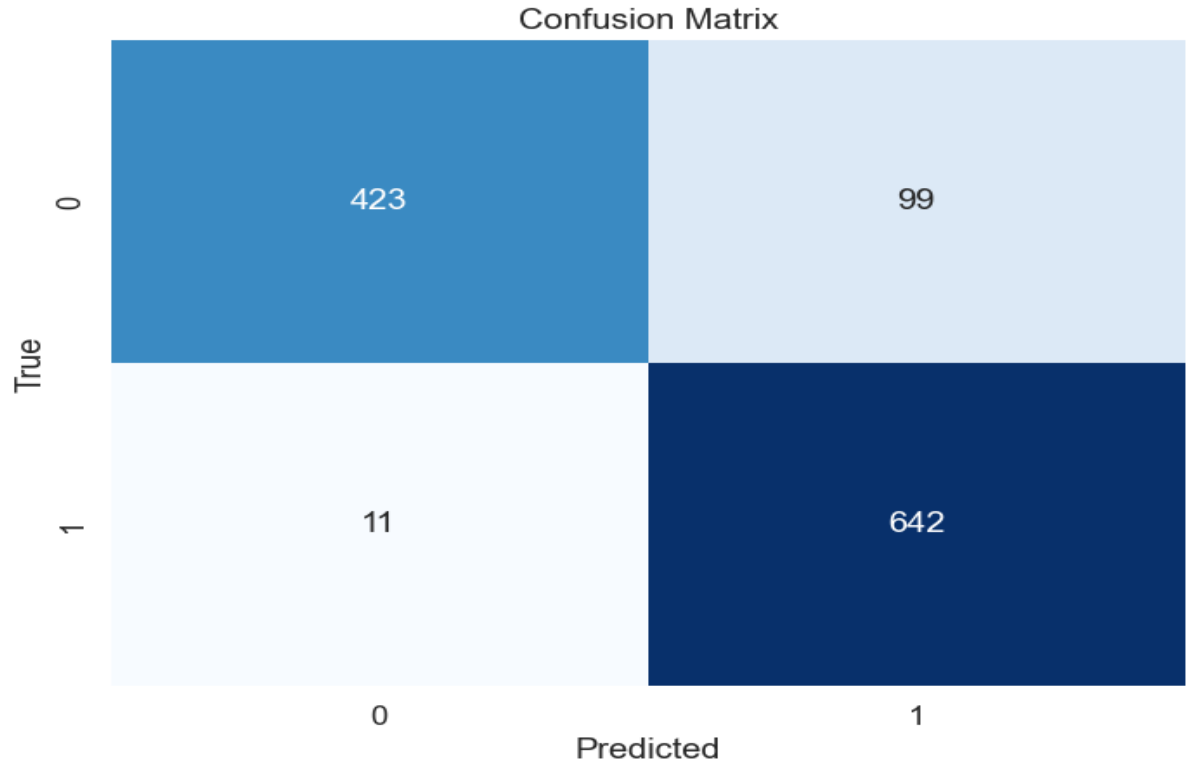


Table 18

```
SVM - Cross-Validation AUC_score: 0.8967497491682949
SVM - Cross-Validation Accuracy: 0.9063829787234042
SVM - Cross-Validation Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.81      0.88       522
           1       0.87      0.98      0.92       653

    accuracy                           0.91      1175
   macro avg       0.92      0.90      0.90      1175
weighted avg       0.91      0.91      0.91      1175

SVM - Cross-Validation F1 Scores: [0.90309817 0.88465298 0.87427702 0.89529725 0.92169874 0.91168478
 0.92121212 0.93872054 0.88481636 0.89252909]
SVM - Cross-Validation Average F1 Score with variance: (0.9027987045231374, 0.0003693551071040771)
```

The SVM model also achieves an AUC score of 0.90 and an accuracy of 0.91, performing well in both precision and recall for both genders. Its F1-scores range between 0.88 to 0.92 with a relatively low variance, suggesting consistent performance. The type I & II errors are not evenly distributed; with a bias towards predicting female.

4.5 Neural Network

Based on scikit-learn (n.d.), The Multi-layer Perceptron (MLP) is a type of artificial neural network capable of learning complex patterns from data. It consists of multiple layers of interconnected nodes (neurons) where each node processes information and passes it to the next layer. Through forward and backward propagation, the MLP adjusts its internal parameters to minimize errors and improve predictions, making it a versatile tool for tasks like classification and regression in machine learning.

Figure 21

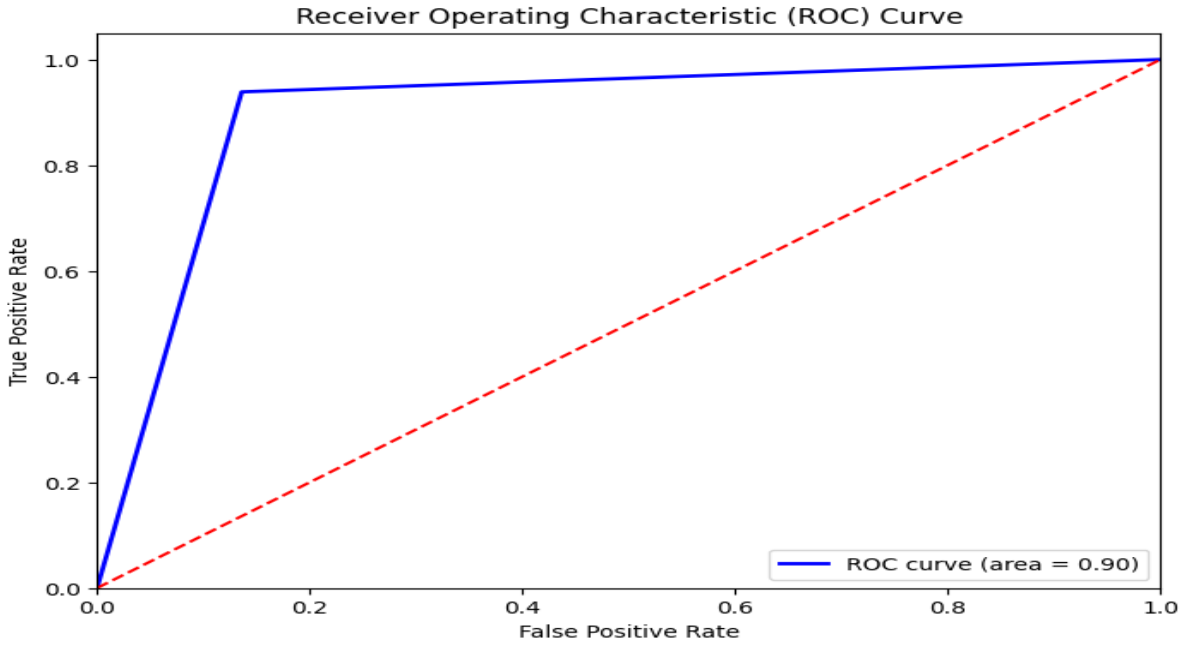


Figure 22

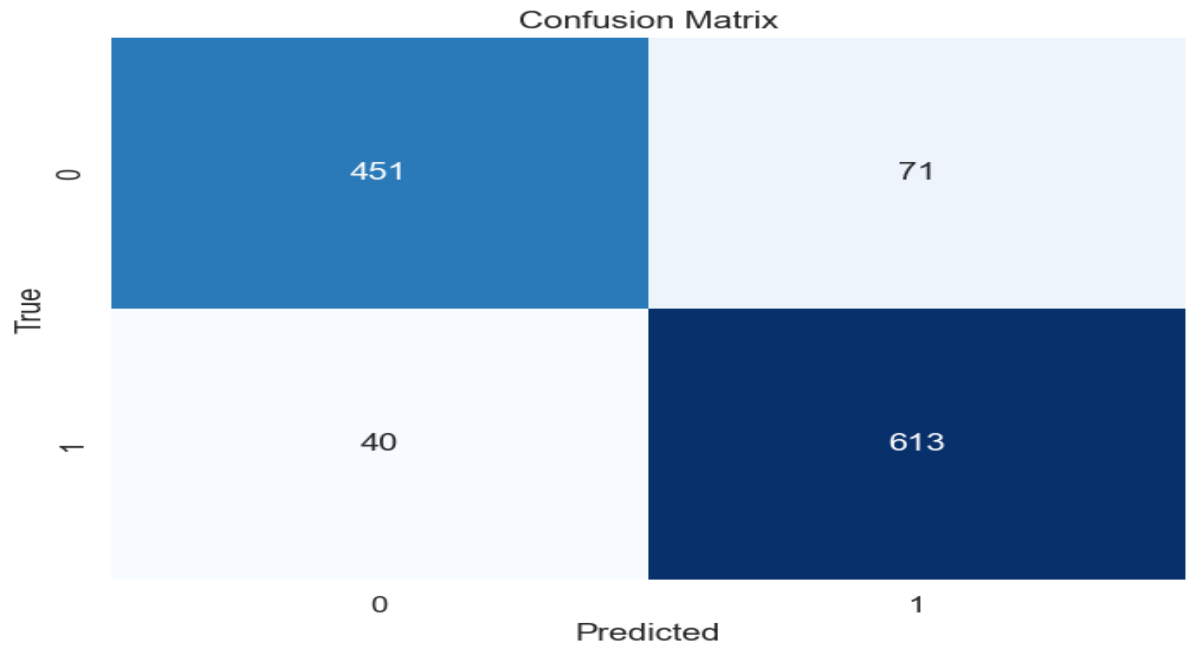


Table 19

```
MLP - Cross-Validation AUC_score: 0.9013644658018107
MLP - Accuracy: 0.9055319148936171
MLP - Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.86      0.89       522
           1       0.90      0.94      0.92       653

    accuracy                           0.91      1175
   macro avg       0.91      0.90      0.90      1175
weighted avg       0.91      0.91      0.91      1175

MLP - Cross-Validation F1 Scores: [0.86690729 0.93994911 0.92211221 0.90524856 0.89685315 0.92161096
 0.9131016  0.89615385 0.89522388 0.87776119]
MLP - Cross-Validation Average F1 Score with variance: (0.9034921795187156, 0.0004266360052678444)
```

The MLP model delivers a good score across all metrics. It shows balanced precision and recall, resulting in a consistence F1-scores with a similar low variance and less bias towards female as SVM.

4.6 Comparison

Figure 23

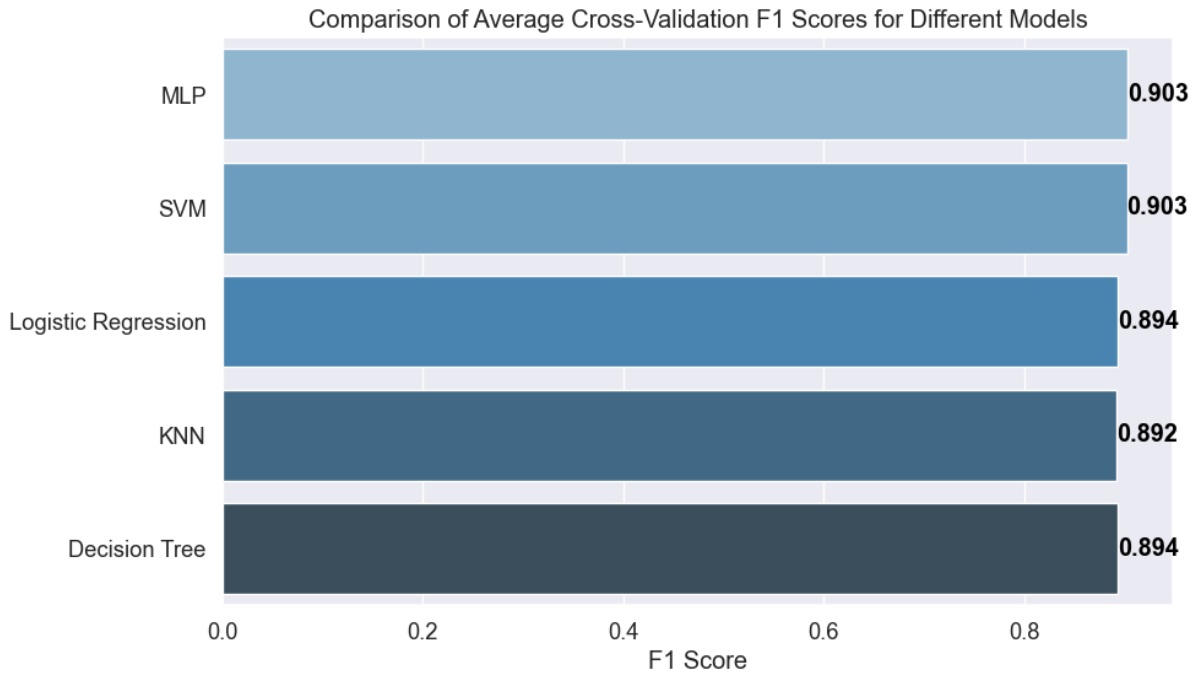


Figure 24

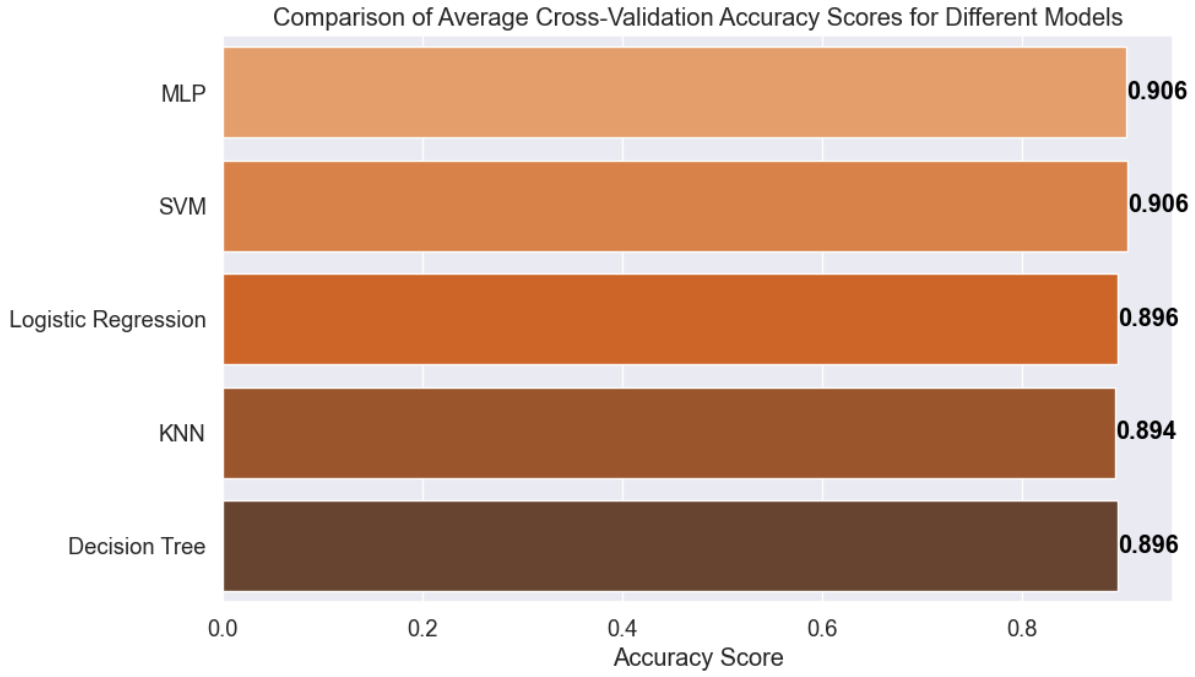
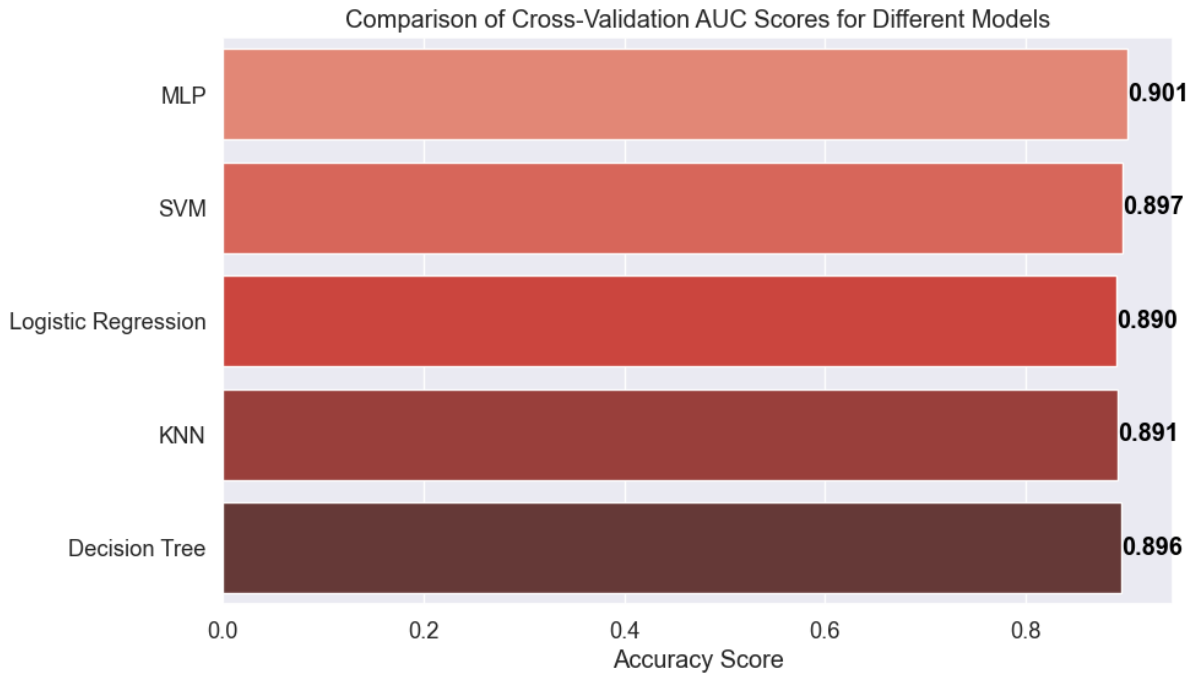


Figure 25



Based on the figures above, all models have the same AUC score and very close in F1 and Accuracy score.

By looking the confusion matrix for all models, the MLP stands out for its consistent and balanced performance. It demonstrates a good overall performance with well-balanced precision and recall for both genders, reflected in F1-score. The low variance in F1 scores signifies stable and reliable predictions. Additionally, the more evenly distributed Type I and Type II errors indicate a well-rounded balance in misclassifications.

The KNN, Logistic regression, SVM and Decision Tree model also performs admirably, showcasing strong metrics overall. However, there's slightly more variability in its predictions, and a slight bias towards predicting females is observed in the error distribution. SVM have the least error in classifying male but most in classifying female. Decision Tree model has the most balance in the error distribution. However, its F1 and accuracy score is less than MLP.

In conclusion, while all models perform reasonably well, the MLP model stands out for its consistent and balanced performance across various evaluation metrics. It showcases stable predictions and a well-rounded approach to classifying genders, making it the most consistent model among those evaluated for this specific task.

5. Regression

5.1 Linear Regression

Figure 26

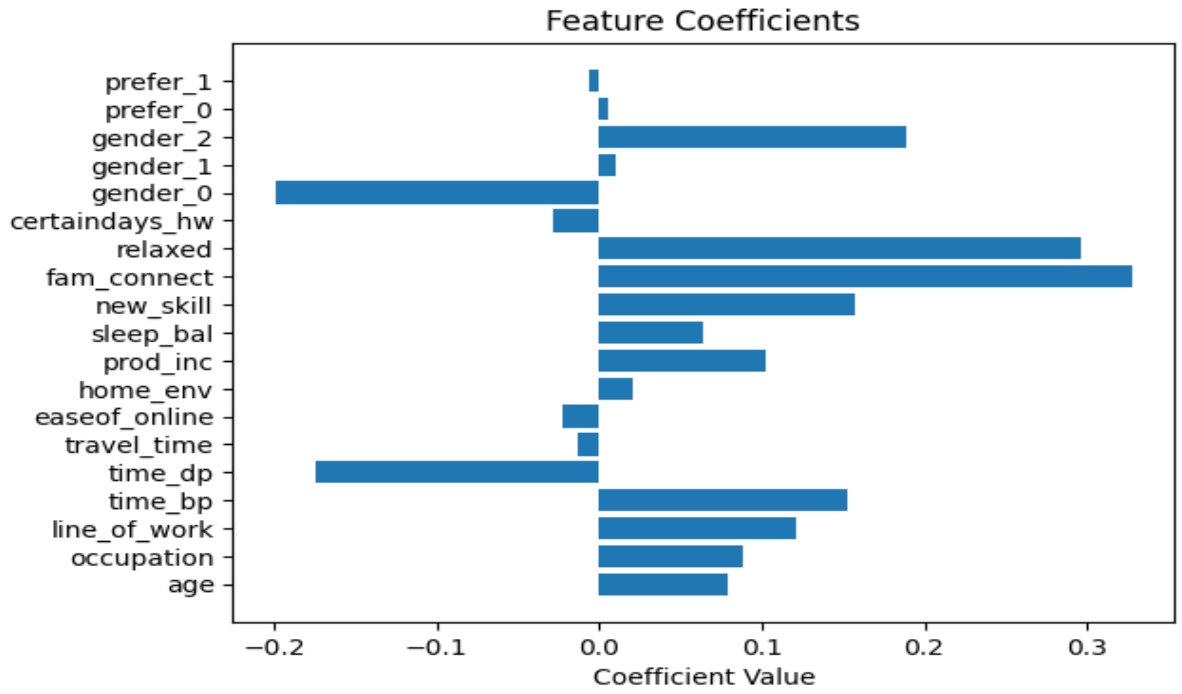


Table 20

```
{'LR - MAE': 0.43790909210397383,
 'LR - R-squared': 0.6594013559393591,
 'LR - Intercept': 0.07980491972650672}
```

The Linear Regression model achieved an MAE of 0.438 and an R-squared value of approximately 0.659. The R-squared value indicates how well the independent variables explain the variance of the dependent variable; in this case, it suggests that around 65.9% of the variability in the 'self_time' can be explained by the model.

5.2 Decision Tree Regression

Based on scikit-learn (n.d.), the Decision Tree Regression in is a predictive model that works by splitting the dataset into smaller subsets based on different features. It constructs a tree-like structure where each internal node represents a feature, each branch represents a decision based on that feature, and each leaf node represents a prediction. By recursively partitioning the data, it creates a flowchart-like model to make predictions for continuous target variables. This model is intuitive, interpretable, and useful for capturing non-linear relationships between features and targets in regression tasks.

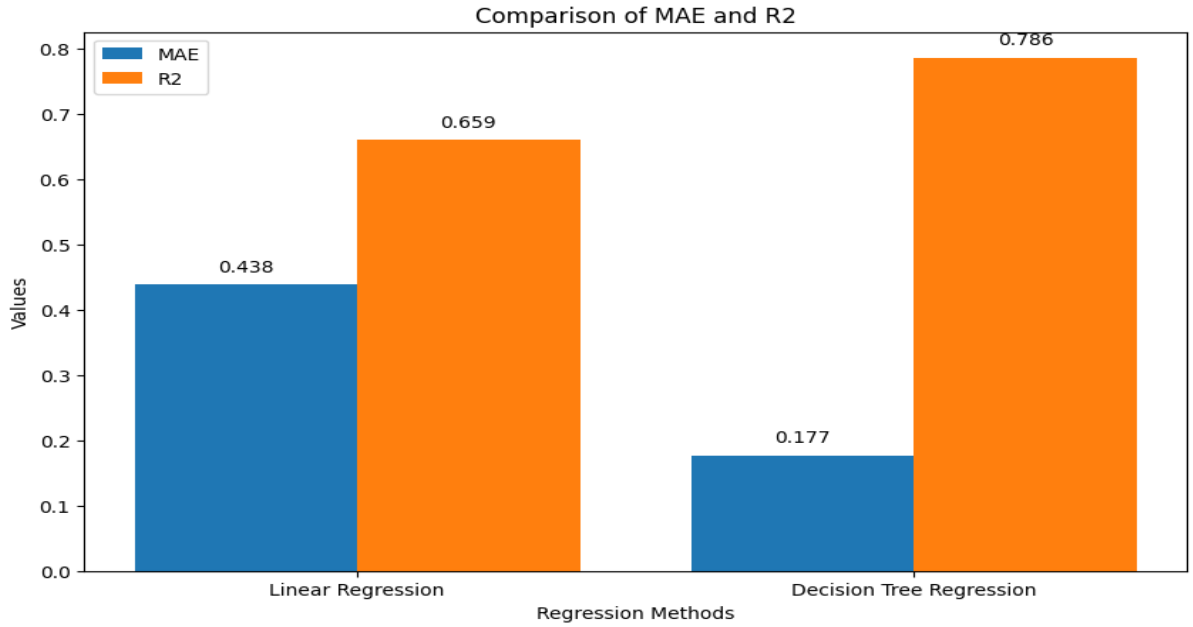
Table 21

```
{'MAE': 0.1769105208442896,
 'R-squared': 0.7855905730986916,
 'Model': DecisionTreeRegressor(random_state=12)}
```

The Decision Tree Regression model resulted in an MAE of 0.177 and an R-squared value of roughly 0.786. This implies that the Decision Tree Regression model, with a higher R-squared value and lower MAE, performed better in predicting 'self_time' compared to the Linear Regression model. The Decision Tree model explained approximately 78.6% of the variability in 'self_time' using the given features

5.3 Comparison

Figure 27



In comparing the two regression models used for predicting 'self_time', the Decision Tree Regression outperformed the Linear Regression model. The Decision Tree model showed a lower MAE compared to the Linear Regression's MAE. Meaning it is less likely that it will predict ‘self_time’ off.

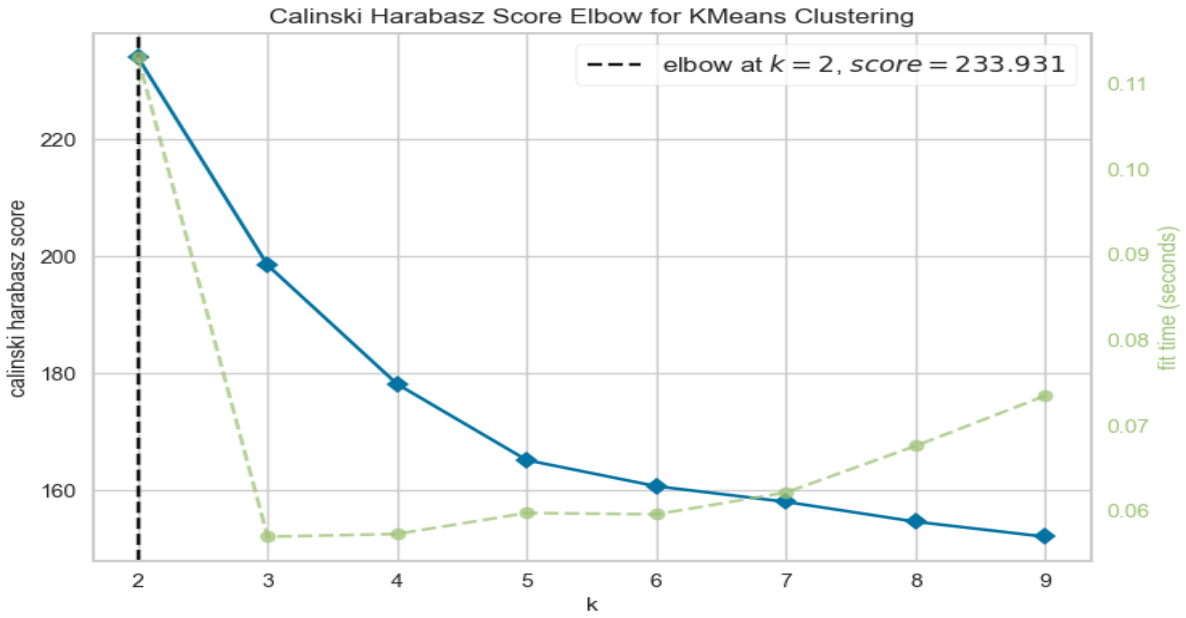
Moreover, the Decision Tree Regression achieved a higher R-squared value, indicating that it explained about 78.6% of the variability in 'self_time'. In contrast, the Linear Regression model had explaining approximately 65.9% of the variability in the dependent variable.

Therefore, the Decision Tree Regression model demonstrated a better predictive accuracy, with a MAE and better explanation of the variability in 'self_time' compared to the Linear Regression model.

6. Clustering

6.1 K-Means

Figure 28



The choice of number of K is chosen to be 3 due to further comparative analysis and because the Calinski-Harabasz score in the figure above does not indicate diminishing returns.

figure 29 indicates the following:

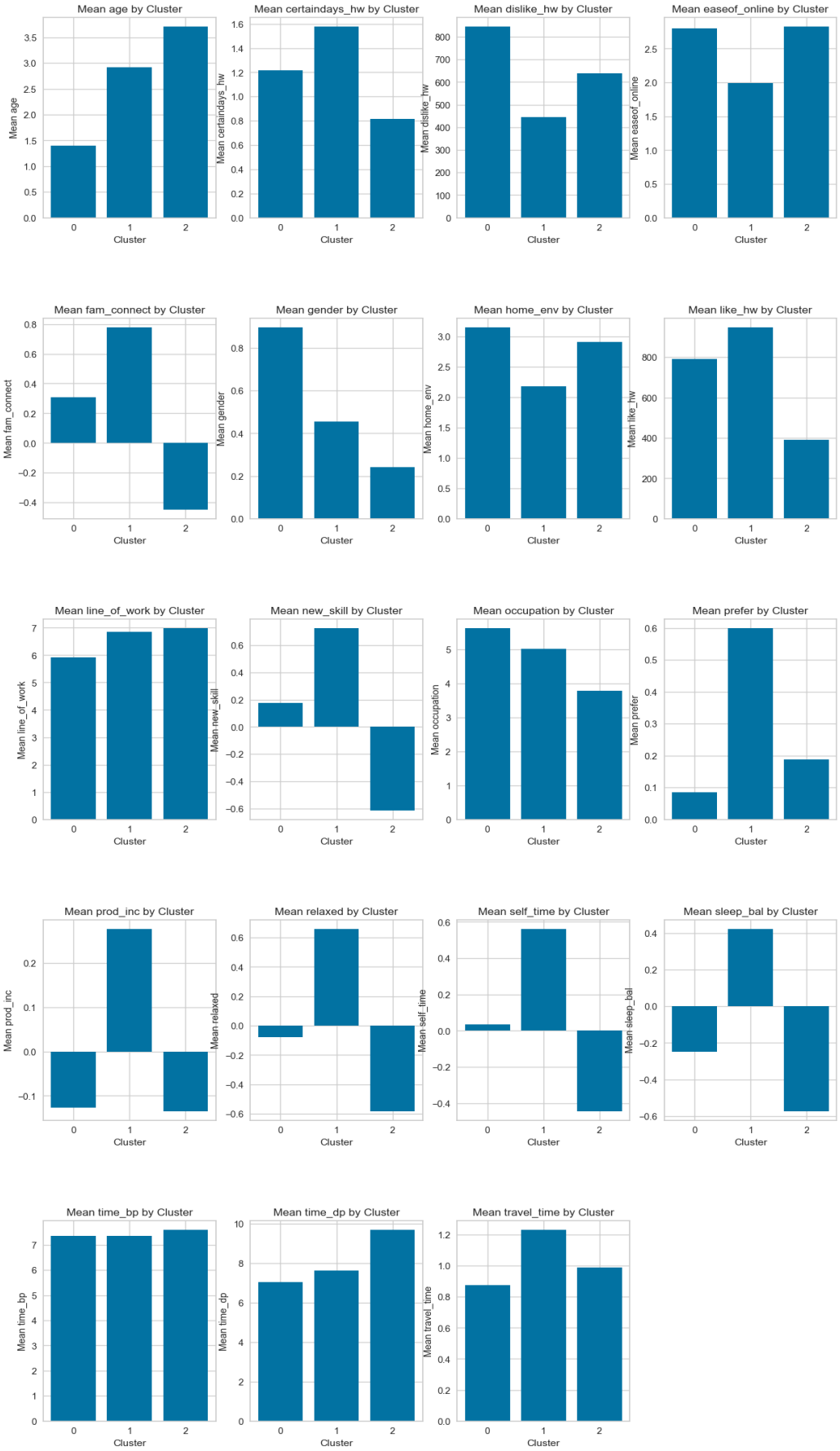
In the first group (Cluster 0), people tend to be younger, mostly male, and work across diverse job sectors. They maintained consistent work hours before the pandemic, with moderately positive feelings about remote work and home environments. However, their productivity, sleep balance, and preference for working from home were more neutral or slightly negative.

The second group (Cluster 1) comprises individuals of moderate to higher age, with a more balanced gender representation, and stable job sectors. They exhibited slightly increased work hours during the pandemic, positive sentiments toward productivity, family connections, and a preference for working from home.

The last group (Cluster 2) consisted of older individuals, primarily female, working in specialized sectors with increased work hours during the pandemic. Their sentiments were more neutral across various factors, with less preference for certain aspects of remote work.

These clusters reveal distinct patterns in terms of time allocation, work environment preferences, and emotional inclinations toward remote work, providing insights into different segments within the dataset.

Figure 29

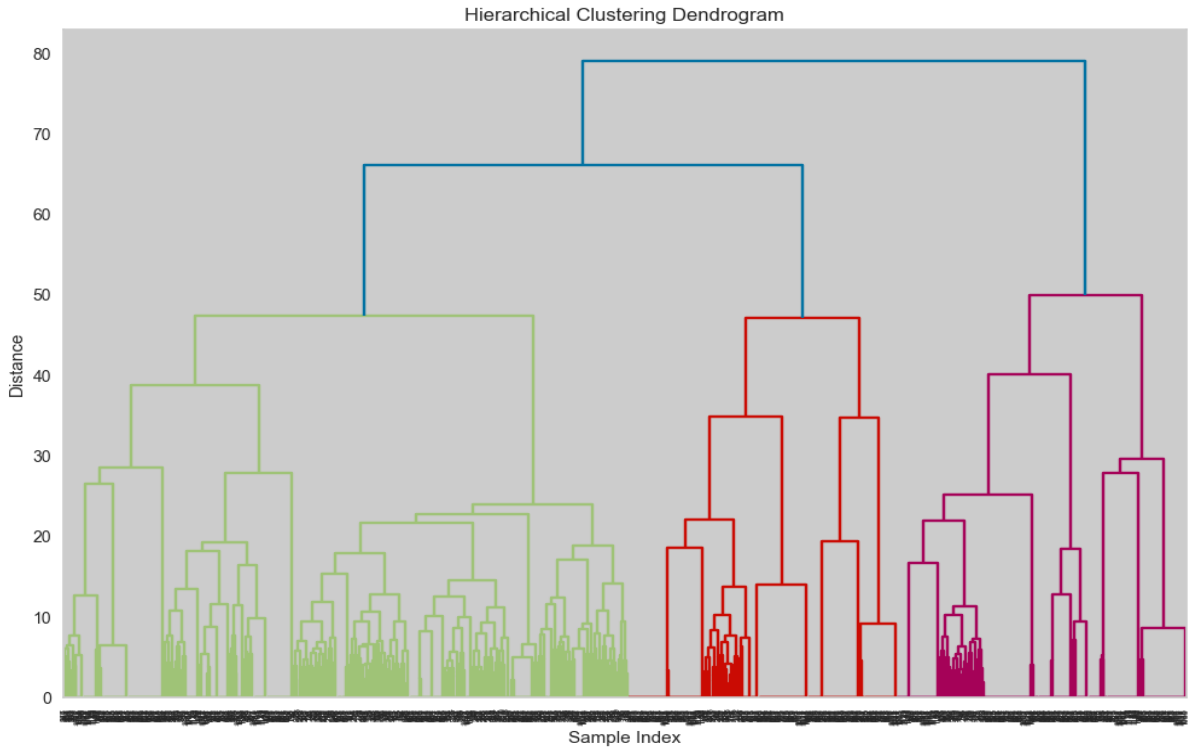

Mean age by Cluster
Mean age
Cluster
Mean certaindays_hw by Cluster
Mean certaindays_hw
Cluster
Mean dislike_hw by Cluster
Mean dislike_hw
Cluster
Mean easeof_online by Cluster
Mean easeof_online
Cluster
Mean fam_connect by Cluster
Mean fam_connect
Cluster
Mean gender by Cluster
Mean gender
Cluster
Mean home_env by Cluster
Mean home_env
Cluster
Mean like_hw by Cluster
Mean like_hw
Cluster
Mean line_of_work by Cluster
Mean line_of_work
Cluster
Mean new_skill by Cluster
Mean new_skill
Cluster
Mean occupation by Cluster
Mean occupation
Cluster
Mean prefer by Cluster
Mean prefer
Cluster
Mean prod_inc by Cluster
Mean prod_inc
Cluster
Mean relaxed by Cluster
Mean relaxed
Cluster
Mean self_time by Cluster
Mean self_time
Cluster
Mean sleep_bal by Cluster
Mean sleep_bal
Cluster
Mean time_bp by Cluster
Mean time_bp
Cluster
Mean time_dp by Cluster
Mean time_dp
Cluster
Mean travel_time by Cluster
Mean travel_time
Cluster

6.2 Hierarchical Clustering

Based on scikit-learn (n.d.), hierarchical clustering is made up of a set of clustering techniques that construct interconnected clusters by combining or dividing them. These clusters form a hierarchical structure and are represented as a tree, where the initial cluster includes all samples at the root, and the individual clusters with only one sample are positioned as the tree's leaves. The number of optimal cluster can be determine visually by using a dendrogram.

Block (2020) describes the mechanism of a dendrogram as a family tree that illustrates how things are related to each other in groups. It's often made after putting things into groups based on their similarities. It is commonly used to figure out the best way to group these things together.

Figure 30



Based on the height of the dendrogram where the three distinct branches fuse the choice of the cluster is 3, indicating a meaningful level of similarity or dissimilarity among the clusters.

From figure 31:

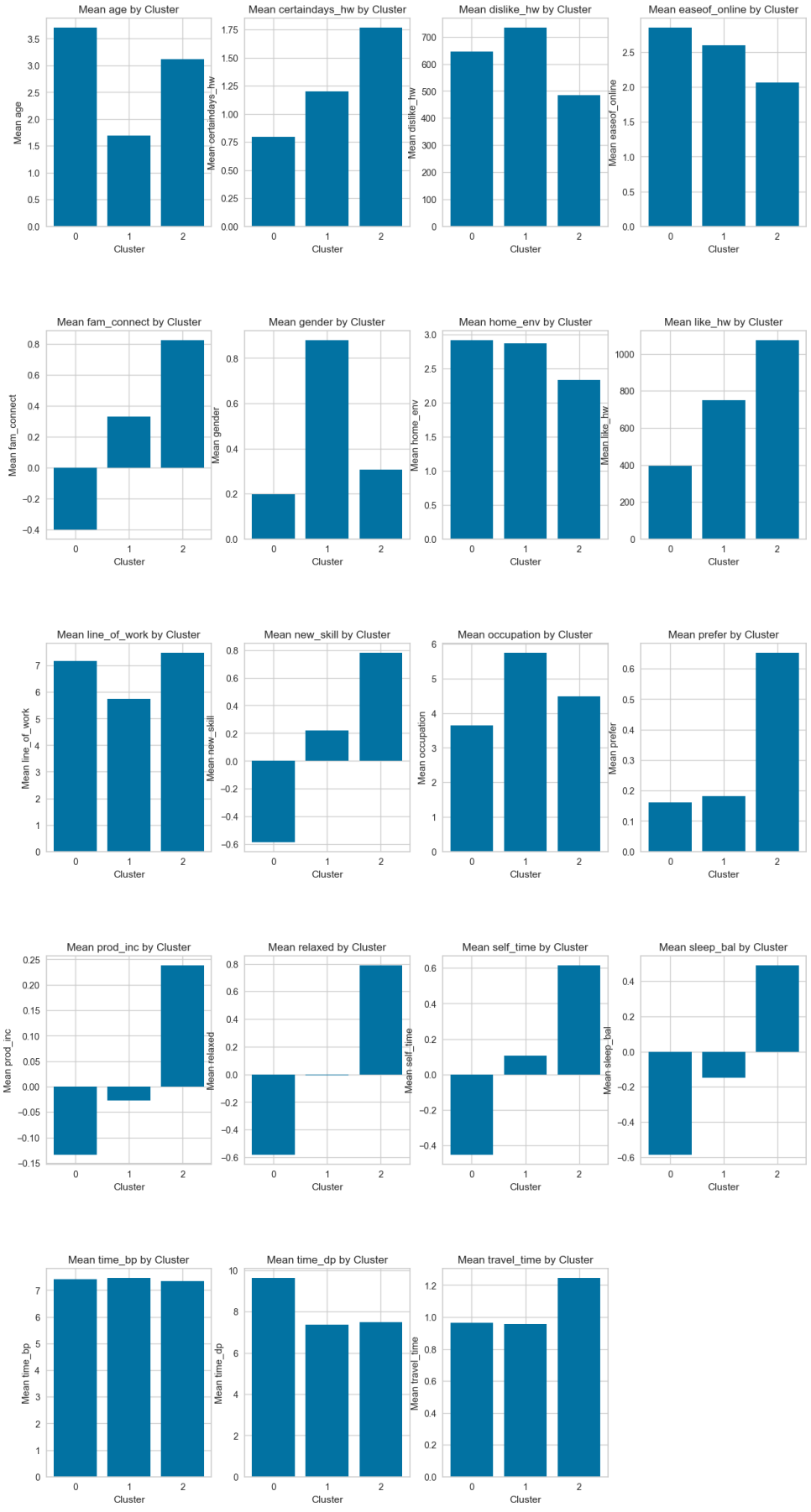
The hierarchical clustering analysis identified three distinct groups based on the collected attributes. In Cluster 0, individuals tend to be older, mostly female, and work in varied occupations across different sectors. They reported spending consistent hours on work both before and during the pandemic, with longer travel times compared to other clusters. Their sentiments towards remote work and home environments were moderately positive, while factors like productivity, sleep balance, and learning new skills leaned towards the neutral or negative side.

In Cluster 1, individuals were generally younger, predominantly male, and engaged in specific job sectors. They have similar work hours before and during the pandemic, with relatively shorter travel times. Their views on remote work and home environments were moderately positive, and they showed mixed sentiments on factors like productivity and family connection.

Lastly, in Cluster 2, individuals were moderately aged, with a more balanced gender representation and worked in sectors with moderate diversity. They reported slightly varied work hours during the pandemic, with moderate travel times. Their perceptions of remote work and home environments leaned more towards positivity, and they showed relatively positive sentiments regarding productivity, family connection, and preferences for certain aspects of remote work.

These clusters highlight diverse profiles in terms of work habits, sentiments, and adaptability during the pandemic across different age groups and work environments.

Figure 31


Mean age by Cluster
Mean age
Mean certaindays_hw by Cluster
Mean certaindays_hw
Mean dislike_hw by Cluster
Mean dislike_hw
Mean easeof_online by Cluster
Mean easeof_online
Mean fam_connect by Cluster
Mean fam_connect
Mean gender by Cluster
Mean gender
Mean home_env by Cluster
Mean home_env
Mean like_hw by Cluster
Mean like_hw
Mean line_of_work by Cluster
Mean line_of_work
Mean new_skill by Cluster
Mean new_skill
Mean occupation by Cluster
Mean occupation
Mean prefer by Cluster
Mean prefer
Mean prod_inc by Cluster
Mean prod_inc
Mean relaxed by Cluster
Mean relaxed
Mean self_time by Cluster
Mean self_time
Mean sleep_bal by Cluster
Mean sleep_bal
Mean time_bp by Cluster
Mean time_bp
Mean time_dp by Cluster
Mean time_dp
Mean travel_time by Cluster
Mean travel_time
Cluster

6.3 Comparison

Figure 32

The hierarchical clustering and K-means analyses provide different perspectives on how individuals are grouped based on various factors related to their work habits and personal preferences during the pandemic.

In hierarchical clustering, Cluster 0 comprises individuals characterized by a higher average age and a longer duration spent working during the pandemic. They display relatively mixed feelings regarding productivity increase, new skill acquisition, and connections with family. Cluster 1, in contrast, represents individuals with a lower average age and varying durations of work during the pandemic. This group shows differences in preferences for working from home and varying feelings of relaxation and self-time. Cluster 2 comprises individuals with a moderate age range and varying durations of work during the pandemic. Their preferences differ concerning productivity increase, liking the home environment, and certain days of remote work.

On the other hand, K-means clustering shows different clusters. Cluster 0 consists of individuals with lower average age, varying durations of work during the pandemic, and mixed preferences for productivity, new skill acquisition, and liking the home environment. Cluster 1 represents individuals with higher average age and varying durations of work during the pandemic, showcasing differences in preferences for productivity, new skill acquisition, and connections with family. Cluster 2 includes individuals with higher average age, notably increased duration of work during the pandemic, differing preferences for productivity, liking working from home, and certain days of remote work.

While both methods aim to group individuals based on similarities, they do so from different angles. Hierarchical clustering emphasises age-related trends and feelings about work and family, while K-means highlights broader preferences and variations in work-related factors among distinct age groups.

References

Block, T. (n.d.). *What is a Dendrogram?.* DISPLAYR. https://www.displayr.com/what-is-dendrogram/

Decision Tree Regression, (n.d.), scikit-learn. https://scikit-learn.org/stable/auto_examples/tree/plot_tree_regression.html

1.17. Neural network models (supervised). (n.d.). scikit-learn. https://scikit-learn.org/stable/modules/neural_networks_supervised.html

Harikrishnan, H. (2023, July). *Psychological Effects of COVID.* Kaggle. https://www.kaggle.com/datasets/hemanthhari/psycological-effects-of-covid/discussion/428480

Hierarchical clustering. (n.d.). scikit-learn. https://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering