PRICE PREDICTION FOR PRIVATE TEACHER PER HOUR

Presentation by Tal Yaakobi and Orel Harazi

INTRODUCTION

מי מאיתנו לא התקשה בלימודי מקצוע מסוים? לשם כך לעיתים אנו פונים למורים פרטיים.
השוק מלא במגוון מורים בשלל מחירים שונים ומגוונים
שמתבססים על משתנים שונים כמו השכלה, ניסיון,
פלטפורמות,שפות לימוד, מיקום וכד'. לעיתים מורים פרטיים
מעריכים את עצמם במחירים מופרזים\נמוכים מהמצופה.
הפרויקט שלנו בא לבדוק האם ניתן לחזות מחיר לשעה
של מורה פרטי ואילו מאפיינים ישפיעו על המחיר.

SCRAPING & CRAWLING

```
#table
table = soup.find("table", class_="strong-boxes")
if table:
    rows = table.find_all("tr")

    #languages
    try:
        for row in rows:
            th = row.find("th")
            if th and th.get_text().strip() == "The class is taught in:":
                td = row.find("td")
                storage["languages"]=len(td.get_text().strip().split(","))
    except:
        storage["languages"]=0

    #levels
    try:
        for row in rows:
            th=row.find("th")
            if th and th.get_text().strip() == "Student level:":
                divs=row.find_all('div')
                levels=list()
                for div in divs:
                    levels.append(div.get_text().strip())
                storage["levels"]=levels
    except:
        storage["levels"]=0
```

ראשית הוצאנו את כל הלינקים של כל המורים בעזרת ספריית requests.

במהלך תהליך ההרכשה עברנו על הנתונים שנמצאים בעמוד של כל מורה והוצאנו אותם דרך זיהוי המאפיינים בקוד הhtml של עמוד של מורה מסוים בעזרת שימוש בBeautifulSoup .

```
try:
    meta_review = soup.find("meta", itemprop="ratingCount")
    storage["review"]=meta_review["content"]
except:
    storage["review"]=0
```

CRAWLING PROBLEMS

Education

Translate this text using Google Translate.

2017 – 2020 : École Supérieure de l'Aéronautique et des Technologies.
Cycle ingénieur en Géomatique et Topographie (Major de promotion 2018 – 2019 – 2020).
2015 – 2017 : École Supérieure de l'Aéronautique et des Technologies.
Cycle préparatoire Maths – Physique (Admise mention Bien).
2015 : Baccalauréat Mathématiques.

```
['"Children\'s music"', 'Ableton live', 'Accordion', 'Act test
r', 'Adobe indesign', 'Adobe photoshop', 'Adobe premiere', 'Af
lgorithms', 'Alternative education', 'Alternative health', 'Anat
'Architecture', 'Argentine tango', 'Armenian', 'Art history', 'A
oduction', 'Audio recording', 'Automotive technology', 'Aviatior
all', 'Bass guitar', 'Belly dance', 'Biochemistry', 'Biology', '
'Biotechnology', 'Bollywood dance', 'Bosnian', 'Botany', 'Boxin
```

1.עמודת ההשכלה מבוססת על טקסט חופשי שהמורה כתב על עצמו. היינו צריכים לדייק את ההרכשה.

2.קיימים מורים שהטקסט החופשי שלהם היה בשפה שונה מאנגלית לכן היינו צריכים לתרגם בעזרת ספריות תרגום.

3.עמודת הskills כוללת מעל ל450 ערכים שונים. בהתחלה יצרנו רשימה לכל מורה עם היכולות שהוא רשם על עצמו.

רמת ההשכלה של המורה והתחום שהוא מלמד הם פקטורים שמשפיעים על תמחור. בתהליך הטיפול בנתונים הבנו ששימוש ברשימות ההשכלה והיכולות של כל מורה בget_dummies יצרו לנו מעל ל500 עמודות בינאריות.

SOLUTION

STEP 1

חלוקה לקטגוריות

STEP 2

יצירת רשימות בהתבסס על ערכים נפוצים

```
word_counts = Counter()
for skills_list in test_df['skills']:
    words = re.findall(r'\w+', skills_list)
    word_counts.update(words)

sorted_word_counts = word_counts.most_common()

for word, count in sorted_word_counts:
    print(f'{word}: {count}')

num_unique_skills = len(word_counts)
print(f"\nTotal number of unique skills: {num_unique_skills}")
```

```
English: 911
Math: 581
Music: 407
French: 377
Spanish: 350
music: 334
Grammar: 311
```

STEP 3

סיווג השכלה ותחום לימודים של מורה לעמודות בינאריות

SCRAPING & CRAWLING

	url	name	location	at_student_home	at_online	at_teacher_home	rate	review	price	audience	...	Teach Humanities	Teach Language Skills
0	https://www.apprentus.com/en/private-lessons/c...	Amin	Morocco	At student's location: Around Casablanca, Morocco	1	0	4.95	34	62.45	[7-12, 13-17, 18-64, 65+]	...	0.0	0.0
1	https://www.apprentus.com/en/private-lessons/i...	Marina	Netherlands	At student's location: Around London, United K...	1	0	4.85	44	111.89	[4-6, 7-12, 13-17, 18-64, 65+]	...	0.0	0.0
2	https://www.apprentus.com/en/private-lessons/h...	Sebastian	Germany	0	1	At teacher's location: 64546 Mörfelden-Walldor...	5	28	223.78	[13-17, 18-64, 65+]	...	0.0	1.0
3	https://www.apprentus.com/en/private-lessons/o...	Haidar	Lebanon	0	1	0	0	0	86.15	[7-12, 13-17, 18-64, 65+]	...	0.0	0.0

לאחר התיקונים הרכשנו מידע על 7000 מורים מהאתר עם 36 פרמטרים.
סה"כ 252,000 נתונים אותם ייצאנו לcsv.
אספנו לינקים מ280 עמודים. בכל עמוד יש 25 קישורים לדפי מורים.

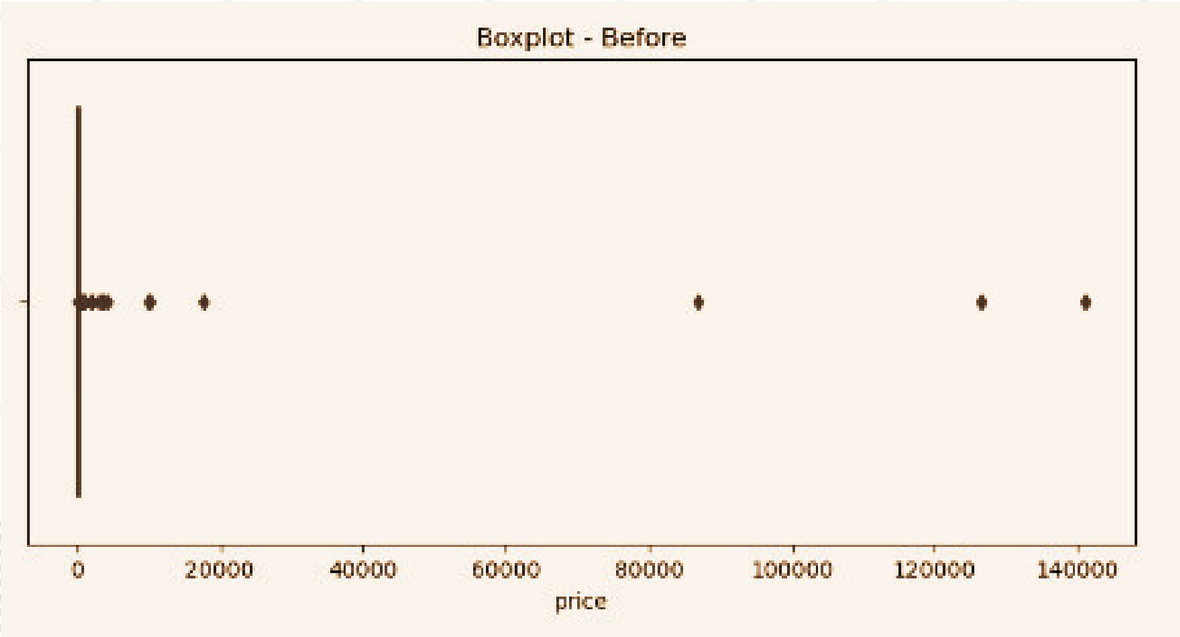
DATA CLEANING

```python
def get_region(country):
    try:
        country_info = CountryInfo(country)
        continent = country_info.region()
        region = country_info.subregion()

        # Customize the region classification based on your requirements
        if continent == 'Africa':
            if region in ['Northern Africa', 'Western Africa']:
                return 'West Africa'
            elif region in ['Eastern Africa', 'Middle Africa', 'Southern Africa']:
                return 'East Africa'
            else:
                return 'Central Africa'

        elif continent == 'Asia':
            if region in ['Eastern Asia', 'South-Eastern Asia']:
                return 'East Asia'
            elif region in ['Southern Asia']:
                return 'South Asia'
            elif region in ['Western Asia', 'Central Asia']:
                return 'West Asia'
            else:
                return 'Central Asia'
```

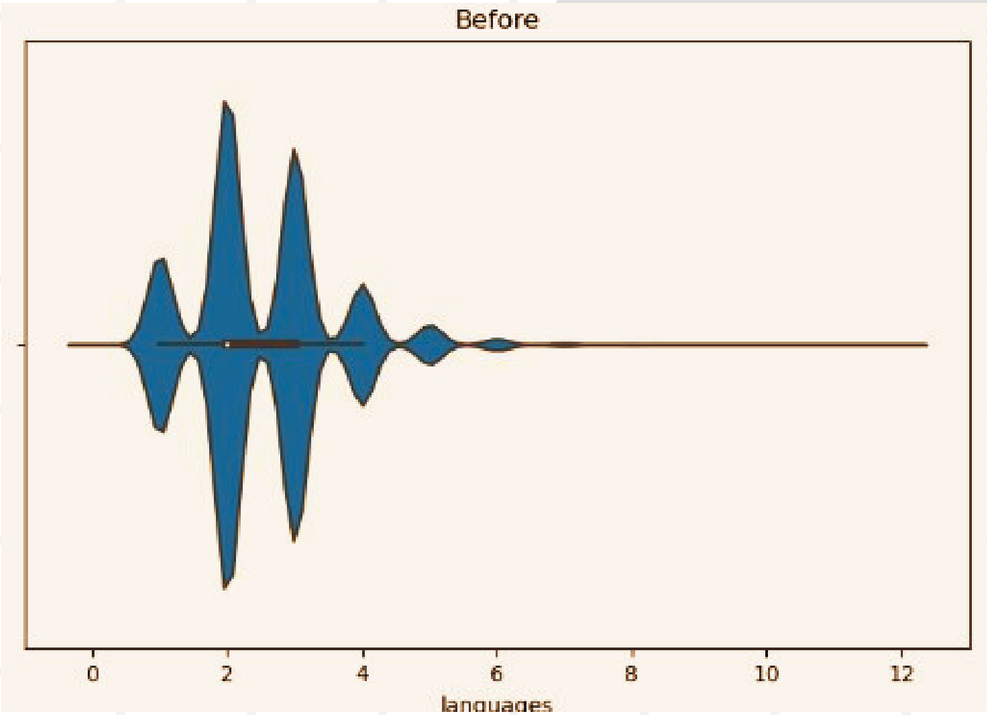
```python
columns_to_replace = ["at_student_home", "at_teacher_home", "trusted_teacher", "at_online"]
for column in columns_to_replace:
    test_df[column] = pd.to_numeric(test_df[column], errors='coerce')
    test_df[column] = test_df[column].apply(lambda x: 1 if x != 0 else x)

current_year = datetime.datetime.now().year
test_df['seniority'] = current_year - test_df['seniority']

for column in ['audience', 'levels']:
    # Split values in each cell and create dummy variables
    values = test_df[column].str.replace('[\[\]]', '', regex=True).str.split(', ')
    encoded_cols = pd.get_dummies(values.apply(pd.Series).stack()).sum(level=0)

    # Update column names to remove brackets and strip extra apostrophes
    encoded_cols.columns = encoded_cols.columns.str.strip("'").str.replace(r"(\d+)-(\d+)", r"\1-\2")

    # Concatenate encoded columns with the original DataFrame
    test_df = pd.concat([test_df, encoded_cols], axis=1)

test_df['region_code'] = pd.factorize(test_df['region'])[0]
test_df['location_code'] = pd.factorize(test_df['location'])[0]
test_df = test_df.drop(columns=['audience', 'levels'])
```

OUTLIERS

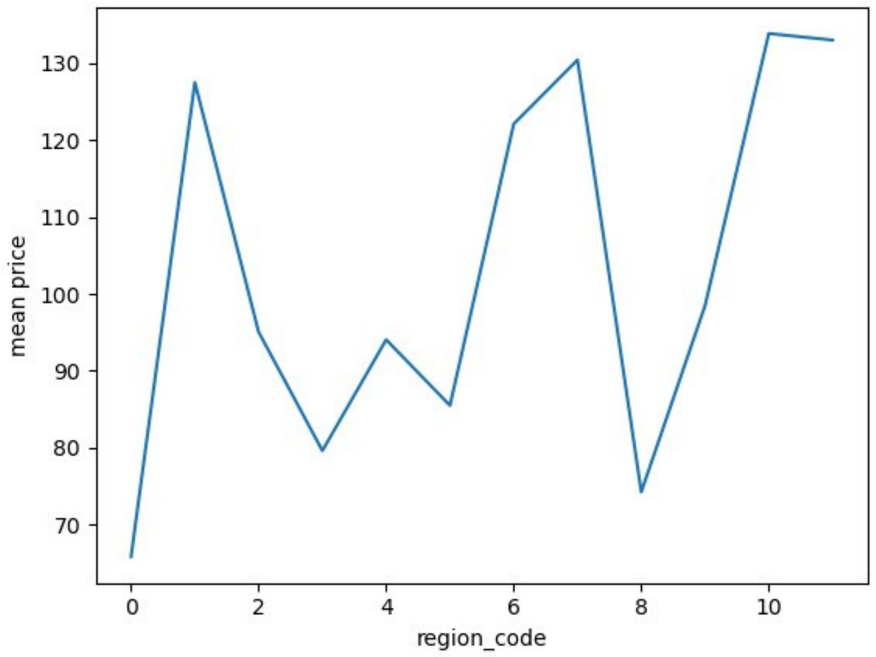
PRICE

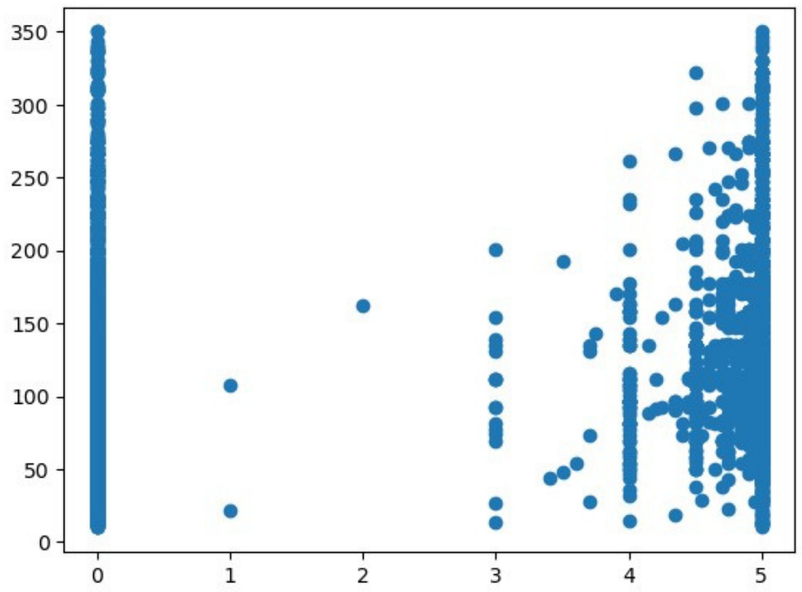


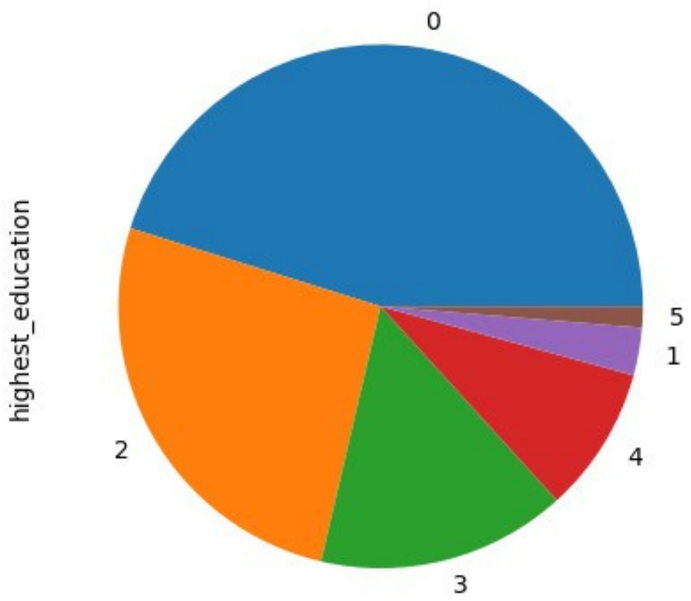
LANGUAGES



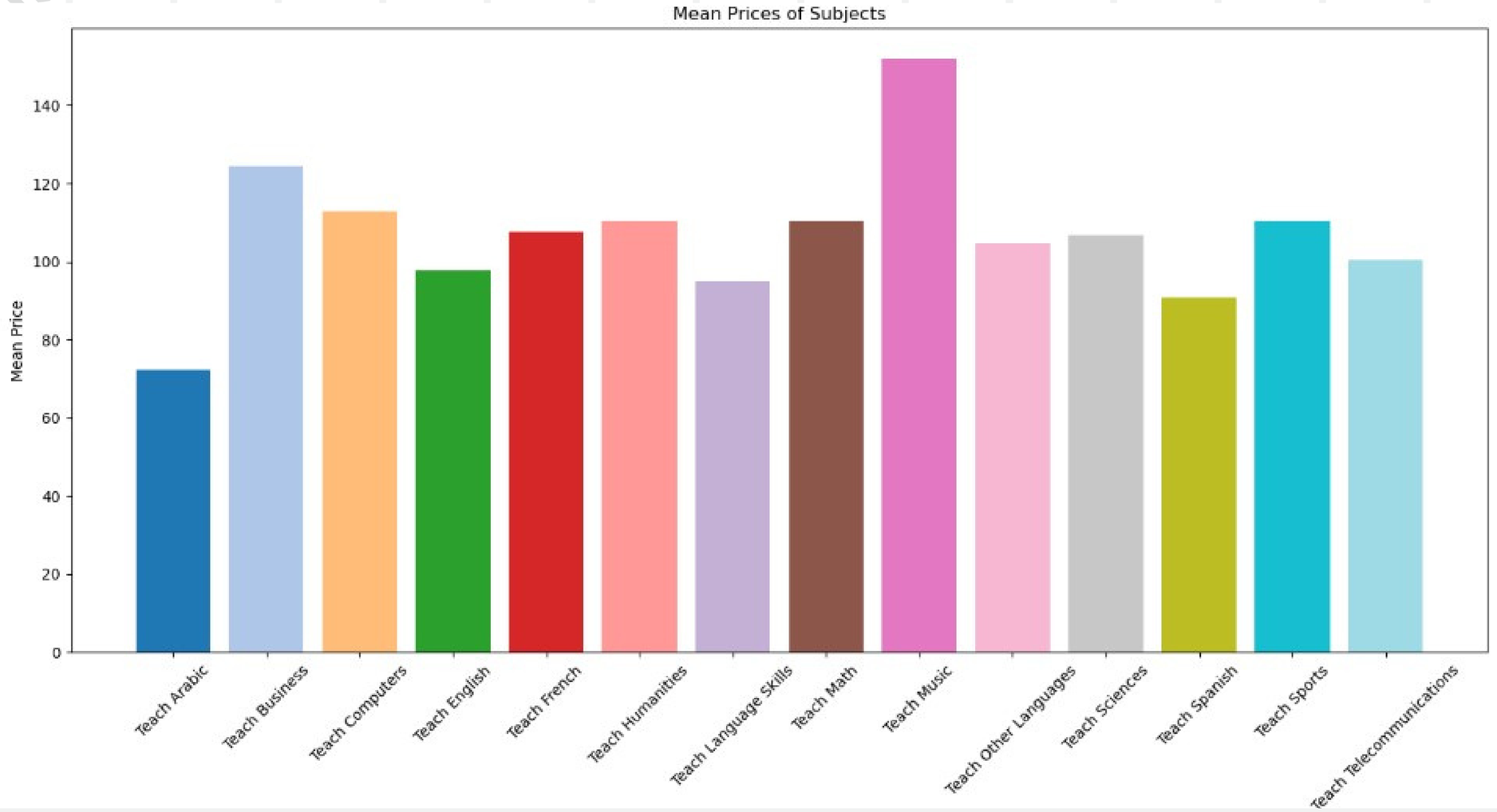
VISUALIZATIONS







VISUALIZATIONS



עשינו מגוון ויזואליזציות שונות בין פרמטרים שונים. שלב זה עזר לנו להבין אילו נתונים ניתן למחוק וכי לא קיים קשר לינארי ברור שיעזור לנו לחזות את הנתונים. שיערנו כי מודל הרגרסיה הליניארית שבחרנו מבעוד מועד לא יספק חיזוי הולם.

MECHINE LERNING

⚙ MODEL 1 - LINEARREGRESSION

```
trained_model_LR,predicted_LR = train_1st_model(X_train, y_train)
predicted_vals_LR = predict_1st(trained_model_LR, X_test)
evaluate_value_LR = evaluate_performance_1st(y_test,predicted_vals_LR)
```

```
print(f"predicted LR:{predicted_LR}\nf1_score:{evaluate_value_LR}")
```

```
predicted LR:0.19286214621850284
f1_score:0.21057684492672435
```

⚙ MODEL 2- RANDOMFORESTREGRESSOR

```
model=RandomForestRegressor(n_estimators=300, random_state=45).fit(X, y)
```

```
print(model.score(X_test, y_test))
```

```
0.9141378753101608
```

CONCLUSION

מטרת הפרויקט הייתה לחזות את המחיר השעתי של מורים פרטיים בהתבסס על גורמים שונים כמו ניסיונם, מיקומם, הנושא והכישורים שלהם. הנתונים שנאספו נותחו באמצעות טכניקות סטטיסטיות, גרפים שונים ולמידת מכונה כדי לזהות את הגורמים המשפיעים על התעריפים השעתיים של מורים פרטיים.
לבסוף הגענו לדיוק של 0.91 בעזרת מודל יער עצי החלטה.

המסקנה אליה הגענו בפרויקט היא:

ניתן לחזות את המחיר השעתי של מורה פרטי על סמך נתונים של מורים אחרים מכל העולם בתחומים שונים.

REFERENCE

APPRENTUS
https://www.apprentus.com/en/private-lessons/online

THANK YOU

Presentation by Tal Yaakobi and Orel Harazi