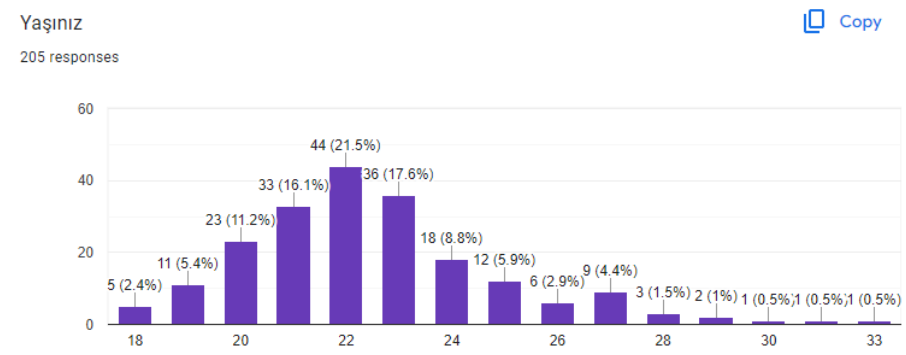
YAPAY ZEKA ÖDEV 2

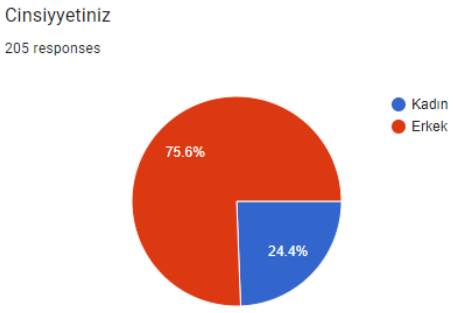
Poll Classification with Cross Validation

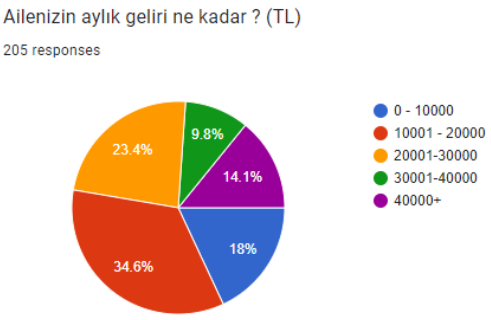
Bakhtiyar Abbasov – 18011114 **Tural Azimov** – 19011903

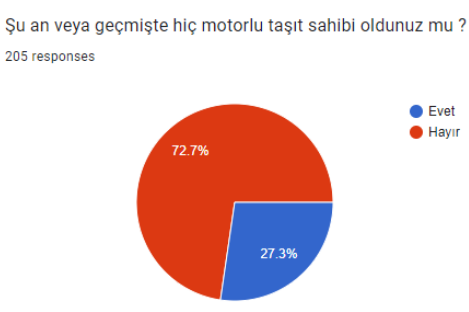
In this project, we aimed to classify our poll results based on the responses to the first nine questions. The initial two questions were of a general nature, asking about the participants' age and gender. Subsequent questions included subtle indicators, such as "time spent in traffic" and "driver license status," which provided insights into whether individuals owned or had previously owned a vehicle. Our primary objective was to determine the presence or absence of vehicle ownership based on the participants' responses to these questions.

To collect our data, we utilized Google Forms and designed a poll consisting of 10 easily answerable questions. Through the collaborative efforts of student groups, friends, and various communities, we have obtained a total of 205 responses as of the present time. Some of the key statistical information regarding our data is depicted in the pie charts and graphs below.









To accurately classify the poll responses, we employed a combination of strategies. Firstly, we utilized multiple classifiers, taking advantage of their unique algorithms and decision-making approaches. By considering the predictions from various classifiers, we aimed to achieve a more comprehensive understanding of the data and improve the accuracy of our results. Additionally, to evaluate the performance of these classifiers, we implemented cross-validation. This approach allowed us to train and test the classifiers on different subsets of the data, providing a more robust assessment of their effectiveness.

To further improve the precision score of our classification model, we employed feature selection techniques, specifically Principal Component Analysis (PCA). By performing PCA, we aimed to reduce the dimensionality of our dataset while retaining the most informative features. As seen in Table 1, this process helped us identify the key variables that had the most significant impact on determining vehicle ownership. Additionally, we employed data normalization techniques to bring the numerical features to a consistent scale, ensuring that no particular feature dominates the classification process due to its larger magnitude. By normalizing the data, we could effectively compare and weigh the different features based on their relative importance, leading to a more precise classification model.

CLASSIFIER NAME	MEAN ACCURACY	
	BEFORE PCA	AFTER PCA
Logistic Regression	0.9085	0.9083
Linear SVC	0.8745	0.9133
K-NN	0.8983	0.884
Random Forest	0.889	0.8842
Gaussian NB	0.8357	0.8983

Table 1. Mean accuracy before and after PCA

As you can see in Images 1 and 2, for most of our classifier pairs, although there are some differences in t-statistics both negatively and positively, p-values are generally greater than 0.05, suggesting there are no significant differences between these classifiers. But in one pair Linear SVC and K-NN, p value is close to the significance level of 0.05, indicating a marginally significant difference.

In summary, based on the p-values, there is no strong evidence to support significant differences in performance between any of the classifier pairs. However, in some cases, there may be a marginal difference that could be further investigated or considered, depending on the specific context and requirements.

```
Paired t-test results between Logistic Regression and Linear SVC:
t-statistic: -0.436941957343603
p-value: 0.6724492937906833


Paired t-test results between Logistic Regression and K-NN:
t-statistic: 1.8268738109952345
p-value: 0.10099563688611873


Paired t-test results between Logistic Regression and Random Forest:
t-statistic: 1.4607032714656467
p-value: 0.17810890880242497


Paired t-test results between Logistic Regression and Gaussian NB:
t-statistic: 1.0268706524561217
p-value: 0.3312877359722881


Paired t-test results between Linear SVC and K-NN:
t-statistic: 2.2489550338934245
p-value: 0.051090410419916625
```

Image 1. First 5 pairs

```
Paired t-test results between Linear SVC and Random Forest:
t-statistic: 1.52057144885361
p-value: 0.16269056892613237


Paired t-test results between Linear SVC and Gaussian NB:
t-statistic: 0.8377195425345904
p-value: 0.4238756944424229


Paired t-test results between K-NN and Random Forest:
t-statistic: -0.013635095969426088
p-value: 0.9894185778613889


Paired t-test results between K-NN and Gaussian NB:
t-statistic: -0.7917922609625099
p-value: 0.44884643173103655


Paired t-test results between Random Forest and Gaussian NB:
t-statistic: -0.648177731947186
p-value: 0.5330558948873165
```

Image 2. Second 5 pairs