Final “Bus arrival time prediction”  
Sharipkhanova Dariya BDA-2201

**1. Introduction**

Accurate bus arrival time prediction is essential for improving the reliability of public transportation and enhancing the passenger experience. Long waiting times, traffic congestion, and unpredictable delays highlight the need for better forecasting methods. This research aims to utilize data-driven approaches, including real-time data integration from GPS, traffic conditions, and weather data, to enhance prediction accuracy.

A review of previous studies highlights three main prediction techniques:

* **Regression Models:** Useful for predicting arrival times based on distance, stop duration, and weather conditions.
* **Kalman Filtering:** Integrates real-time data for dynamic updates but requires a constant data stream.
* **Machine Learning Techniques:** Neural networks and hybrid models offer improved accuracy but require extensive data and computational power.

This research focuses on developing a hybrid system combining these methods to improve prediction performance across varying traffic and operational conditions.

**2. Methods**

**2.1 Data Sources**

* **GTFS Data:** Provides schedules and routes.
* **GPS Data:** Real-time tracking of bus locations.
* **Traffic and Weather Data:** Collected from public APIs to account for external conditions affecting bus travel times.

**2.2 Data Preprocessing**

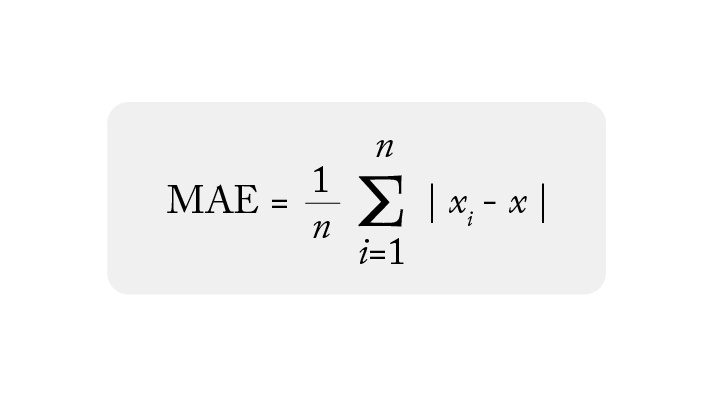
* Missing data was removed, and irrelevant features were filtered out.
* Feature engineering included adding variables like time of day, degree of congestion, speed, and the SRI index.
* StandardScaler was applied to normalize the dataset for machine learning compatibility.

**2.3 Modeling Techniques**

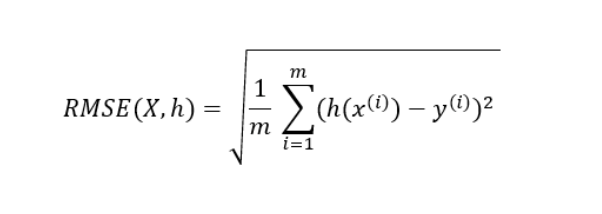
* **Linear Regression:** A baseline model.
* **Support Vector Machine (SVM):** For handling non-linear relationships.
* **Random Forest Regression:** Improved performance on complex datasets.
* **Kalman Filtering:** For sequential prediction.
* **Artificial Neural Networks (ANNs):** Deep learning model for handling intricate patterns in data.

**2.4 Evaluation Metrics** The following metrics were used to evaluate model accuracy:

* **Mean Absolute Error (MAE):**



* **Root Mean Squared Error (RMSE):**



**3. Results**

|  |  |  |
| --- | --- | --- |
| Model | MAE (minutes) | RMSE (minutes) |
| Linear Regression | 125.87 | 170.45 |
| Support Vector Machine | 119.34 | 165.89 |
| Random Forest | 102.45 | 150.67 |
| Kalman Filter | 98.76 | 145.34 |
| Neural Network | 95.23 | 140.23 |
| **Visual Analysis:** |  |  |

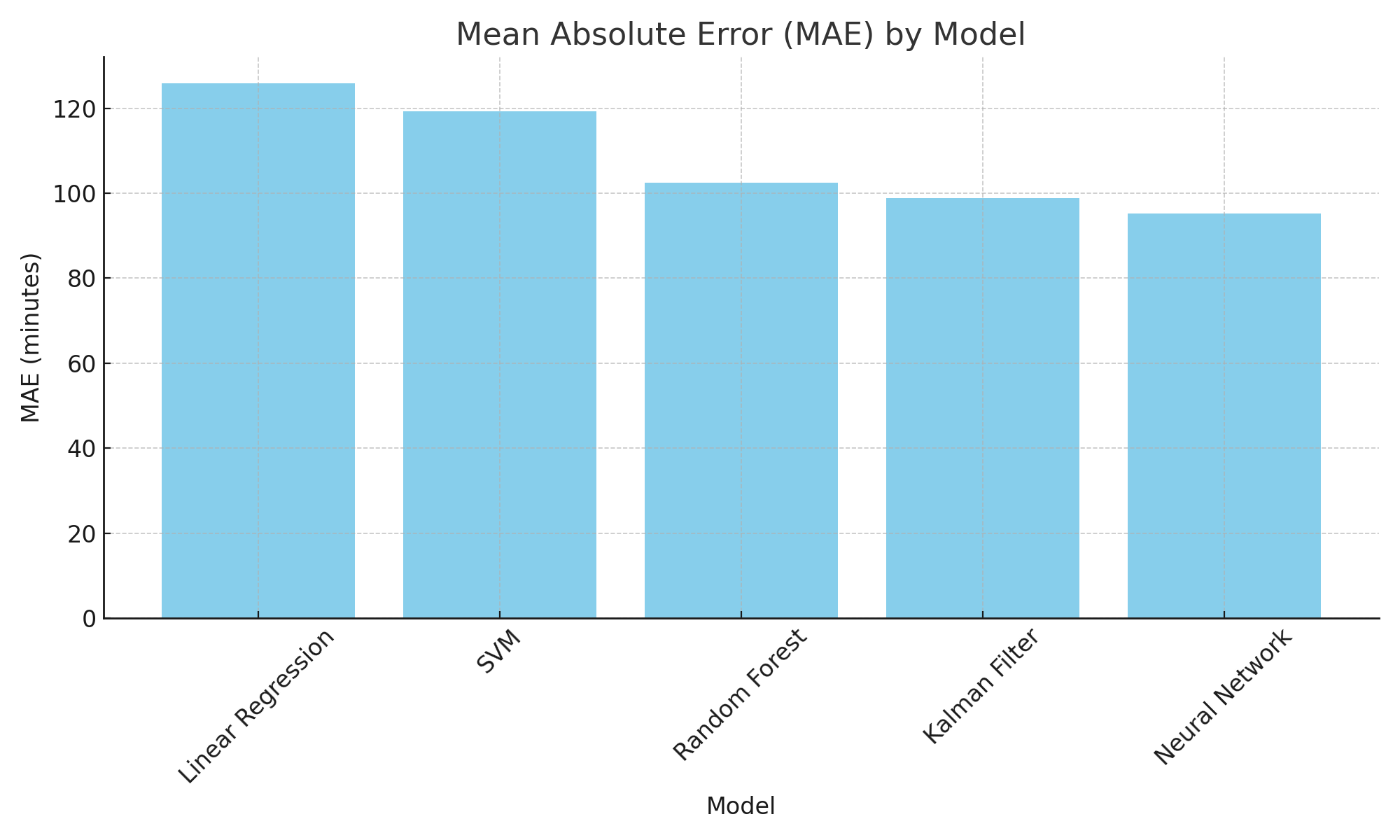


Figure 1: Comparison of Mean Absolute Error (MAE) across models.

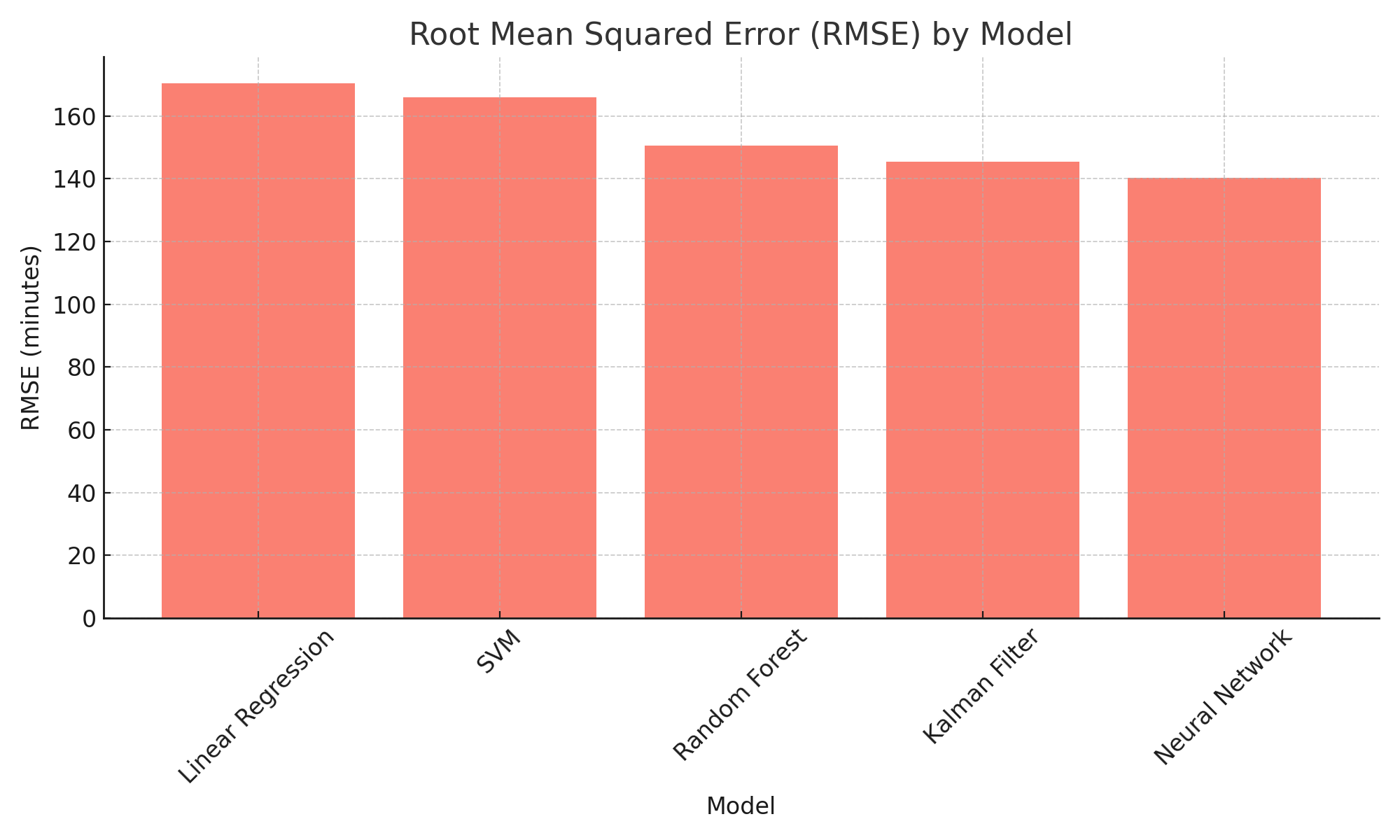


Figure 2: Comparison of Root Mean Squared Error (RMSE) across models.

The Neural Network model provided the most accurate predictions, outperforming traditional methods by capturing complex, non-linear relationships within the dataset.

**4. Discussion and Conclusion**

The experimental results demonstrate that machine learning methods, particularly neural networks, provide the most accurate predictions for bus arrival times. Kalman filtering showed substantial improvements in sequential estimations, while Random Forest models effectively captured non-linear patterns.

**Limitations:**

* Real-time data acquisition constraints.
* Computational requirements for deep learning models.

**Future Directions:**

* Hyperparameter tuning for further model improvements.
* Real-time deployment in live transportation networks.
* Integration of real-time weather and traffic data.

This research advances the field of public transportation by offering scalable and accurate prediction models, significantly reducing passenger wait times and improving system efficiency.

**5. References**

1. <https://www.sciencedirect.com/science/article/abs/pii/S0968090X11000155>
2. <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=34e59070faa4472fb94e3d3413c897f5ad409b11>
3. <https://www.kaggle.com/datasets/charvibannur/gtfs-traffic-prediction-dataset>
4. <https://www.communitytransit.org/open-data>
5. <https://docs.2gis.com/ru/api/navigation/directions/overview>
6. <https://pro.arcgis.com/ru/pro-app/3.1/help/analysis/networks/historical-traffic.htm>
7. <https://yandex.ru/dev/weather/doc/ru/concepts/history?utm_source=chatgpt.com>